

USE OF TEXT MINING TO UNDERSTAND  
REAL ESTATE TRENDS AND  
MARKET DISCUSSION ON SOCIAL MEDIA

A Thesis

Presented to the Faculty of the Graduate School  
of Cornell University

In Partial Fulfillment of Requirements for the Degree of  
Master of Science

by

Hao Rong

August 2018

© 2018 Hao Rong

# **ABSTRACT**

The housing bubble is one of the most urgent social problems to address in China. To guide healthy investment behavior and make effective regulatory policy, it is essential to understand how real estate market discussion shifts correspond to the changes in market conditions and social values. By understanding market discussion, not only can we evaluate the efficiency of the current policy, but we can also make better policy decisions in the future. Since most of the market discussion from certain individuals or organizations are posted online in the form of articles, text mining could be a potent tool in extracting information in order to better comprehend public opinions. This research focuses on obtaining valuable information from text data in social media, organizing and structuring text data, and making convincing statistical inferences on the relationship between online discussions and the actual situation of the real estate market.

## **BIOGRAPHICAL SKETCH**

Hao Rong is an enthusiastic explorer interested in the interplay of Social Science and Data Science, especially in applying and creating methods inspired by data to enhance aspects of urban and social development. He received a Bachelor of Engineering in Urban Planning from South China University of Technology. In the MS Program of Regional Science at Cornell University, his major field of study is Urban and Regional Economy while his minor field of study is Information Systems. After his graduation from the Regional Science program, he will pursue his second master's in Operational Research & Information Engineering at Cornell Tech.

This research is dedicated to all lifelong learners and interdisciplinary researchers.

# ACKNOWLEDGEMENTS

I want to thank my thesis advisors, Professor Kieran Donaghy and Professor David Mimno for their full support of my research interests. They have given me so many useful suggestions for my study and academic career. Additionally, I want to thank Ms. Françoise Vermeylen, the director of Cornell Statistical Consulting Unit for her advice in statistical analysis.

I also want to thank the graduate school of Cornell University and the Regional Science program for giving me great flexibility to explore my interests and future career directions, which I could not imagine to have in any university in China.

Finally, I want to express my gratitude to my boyfriend Braulio Castillo, my family, and my friends for all the love and support. I am grateful for having all of you on my side during my wonderful two years at Cornell.

# TABLE OF CONTENTS

BIOGRAPHICAL SKETCH .....	iii
ACKNOWLEDGEMENTS .....	v
TABLE OF CONTENTS .....	vi
LIST OF FIGURES .....	vii
LIST OF TABLES .....	viii
1 Background.....	1
1.1 Land Finance and Motivation behind the Home-purchasing .....	1
1.2 Regulation and Stimulation Policy in the history .....	4
1.3 Social Media Penetration .....	6
2 Research Objective .....	7
3 Concepts and Tools of Text Mining.....	8
3.1 Text Preprocessing and Representation: .....	8
3.2 Topic Modeling:.....	9
3.3 KeyGraph: .....	10
4 Data Preparation .....	11
4.1 Document Source and Data Wrangling: .....	11
4.2 Topic Modeling:.....	15
4.3 Construct KeyGraph and Clustering Topics: .....	20
5 Regression Analysis - Static Model .....	27
5.1 Summary Statistics on Topics and Main Topics.....	27
5.2 Summary Statistics on Topics and sub Topics.....	30
5.3 Real Estate Statistics .....	36
5.4 Association between Real Estate Statistics and Topics .....	38
5.5 Association between Control Policy and Topics .....	46
6 Time Series Analysis - Granger Causality Test .....	48
6.1 Test Hypothesis.....	49
6.2 Data Preprocessing.....	50
6.3 Result Analysis.....	51
7 Conclusion.....	53
BIBLIOGRAPHY .....	56
APPENDIX.....	58

# LIST OF FIGURES

Figure 4.1.1 Number of Articles Scraped per Month .....	13
Figure 4.1.2 Total Length of Articles per Month (in Chinese Token).....	14
Figure 4.1.3 Average Length of Articles per Month (in Chinese token).....	14
Figure 4.2.1 Elbow Method for Determining Appropriate Amount of Topics.....	16
Figure 4.3.1 Clustering KeyGraph with Nine Main Topics.....	26
Figure 5.1.1 Average Contribution Percentage of Main Topics per Month (2014-06~2017-10).	28
Figure 5.1.2 Correlation of contribution percentage of main topics.....	29
Figure 5.2.1 Average Contribution Percentage of ‘Complaint’ Sub-Topics per Month .....	31
Figure 5.2.2 Average Contribution Percentage of ‘Support’ Sub-Topics per Month.....	32
Figure 5.2.3 Average Contribution Percentage of ‘Life Story’ Sub-Topics per Month.....	32
Figure 5.2.4 Average Contribution Percentage of ‘Investment’ Sub-Topics per Month.....	33
Figure 5.2.5 Average Contribution Percentage of ‘Economic Analysis’ Sub-Topics per Month	33
Figure 5.2.6 Average Contribution Percentage of ‘Regional Price’ Sub-Topics per Month.....	34
Figure 5.2.7 Average Contribution Percentage of ‘Listing’ Sub-Topics per Month.....	34
Figure 5.2.8 Average Contribution Percentage of ‘Policy’ Sub-Topics per Month .....	35
Figure 5.2.9 Average Contribution Percentage of ‘City Service’ Sub-Topics per Month .....	35
Figure 5.3.1 Real Estate Investment and Price Index per Month (2014-06~2017-10).....	37



## LIST OF TABLES

Table 1.2.1 Real Estate Control Policy (1997 ~ 2016) .....	6
Table 4.2.1 Example of Topics and Words with High Probability.....	17
Table 4.2.2 List of Trained Topics and Corresponding Names.....	19
Table 4.2.3 Example of Document and Topics with High Probability.....	20
Table 4.3.1 Different Thresholds and Corresponding KeyGraph.....	23
Table 4.3.2 Different Thresholds and Corresponding Topic Properties.....	24
Table 4.3.3 Naming Explanation of Clustered Topics (Main Topic) .....	25
Table 5.3.1 Variable Explanations.....	38
Table 5.4.1 Regression Analysis on IREA .....	40
Table 5.4.2 Regression Analysis on HCHPI.....	42
Table 5.4.3 Regression Analysis on IRHB90A .....	44
Table 5.5.1 Regression Analysis and Emmeans Analysis .....	47
Table 6.3.1 Granger Causality Test Result .....	52

# **1 Background**

## **1.1 Land Finance and Motivation behind the Home-purchasing**

Spreading real estate risk, especially commercially residential property, has concerned every family in China. Property price in most big cities is so high that an ordinary citizen who does not own a house cannot afford to live in the big cities. In the meantime, population growth is not high enough to fill in the extra houses built in the new towns. With soaring real estate prices absorbing most of the investment, manufacturers and other industries have been suffering from low investment and low interest in entrepreneurship. Local government, however, has been financially relying on land sales since the beginning of 21st century. (Du & Zhang, 2015) Without smoothly transforming the tax policy from indirect tax on land-sell to direct tax, the Chinese government cannot easily restrain the real estate price which is positively related to land price.

To prevent a bubble burst causing a potential financial crisis, the Chinese government has been putting a long-term effort in renewing the regulatory policy to control the real estate price from soaring drastically but also in holding the price within a reasonably small increasing rate. The property market has been experiencing more than four rounds of ups and downs due to recurrent loosening and tightening policies since 2005. The most recent deflation period happened in 2015. In 2016, the real estate market in China experienced its most astonishing price boost in recent history. The price of real estate had almost doubled in some of the major cities. On October 18, 2017, president Xi said "Houses are built to be inhabited, not for speculation (trade for profit)" at the 19th Party Congress in Beijing. "At such an important occasion, it is quite unusual for top leaders to be so straightforward" commented Larry Hu, who thinks this is a signal that the Chinese government will build up a long-term mechanism to cool down the housing market.

Along with the urbanization process of China, the government has been encouraging home purchasing behavior in the cities. This is not only because it can greatly stimulate the urbanization process of population and assets, but because it also provides a significant amount of the government's fiscal revenue. (Wan & Ye, 2016) Compared to a country like the USA that mainly relies on the fiscal revenue on taxes, China has instead been relying on land finances for a very long time, which refers to the fiscal revenue strategy relying in land use processes. The land finance is composed of two main parts, non-tax land revenues and tax land revenues. (Huang, & Cai, 2013) The non-tax revenues mainly refer to land transfer fee, land rent, and paid use fee for the new construction land. While the tax revenues mainly refer to farmland occupancy tax, this can also refer to business tax on sales of real estate and construction, and land value-added tax.

First developed in the early 1990s, the Chinese government has been dependent on this source of fiscal revenue. According to the "China Land Policy Reform" research group, between 2003 and 2006 in Jiangsu province, Zhejiang province and Guangdong province, the tax revenues accounted for 40% of the budgetary revenue while the non-tax revenues accounted for 60% of the extra budgetary revenue. (Tian, & Zhu, 2016) In the tax division system, the land transfer fee is divided into local governmental revenues. Initially, this division system was set to reallocate the fiscal pressure from the central government to the local governments by distributing revenue sources. (Qun, Yongle, & Siqu, 2015) It became, however, the largest portion of local fiscal revenues. According to "China Financial Yearbook 2001-2012", the land transfer fee composed 35% of the total local revenue in 2001 while it composed 69.43% of the total local revenue in 2010. In 2012, the revenue from the local government took up 52% of the total revenue. Meanwhile the proportion of local government revenue rose to 85% of the total cost. Faced with a fiscal deficit of

33%, local governments have no choice but to rely on the land transfer fee of the extra budgetary revenues. (Cao, Feng & Tao, 2008)

Despite the huge dependency on revenue from land transfer fees, the Chinese government has created an upwards increasing spiral where the revenue can be devoted to support urban infrastructure to further enhance the process of urbanization. One of the most important features of the city is that, compared to rural areas, it can provide public services that are not naturally available. (Wan & Ye, 2016) Under this system of an upwards increasing spiral, the revenue from the urban real estate will be transferred into the construction of urban services. Then the enhanced urban service would encourage people to invest and move into the city. In terms of all kinds of urban infrastructure, public services usually require large-scale fixed cost which is why most of these projects are led by the government, since the private sector can hardly yield profit from such long-term projects. Traditionally, a one-time investment is obtained mainly through the accumulation of surplus in the past. This practice greatly inhibits the investment in infrastructure and the development of cities. (Zhao, & Webster, 2011)

An important factor that makes this development model successful is that the Chinese government is rooted in a strong sense of national credits related to growth. (Cao, Feng & Tao, 2008) The planned economy left land as a huge growing credit source. The essence of China's land revenue is to finance the one-time investment through land for the future. The essence of property right is to purchase urban services that are guaranteed in the future. Each city government can be seen as a company that issues stock of the city. Therefore, part of the home-purchasing behavior can be interpreted as an investment choice even if the demand results from living, because every homeowner needs to consider how much their home will be evaluated in the future, which is theoretically close to how much credit that local government has in building service in the future.

Although it is commonly acknowledged that there are bubbles in China's housing market, it does not interfere with the opinion that the spillover payment for housing is partially explained by future services. (Wang, Hou & He, 2017) This also explained the confusion over the stock market being sluggish in China since the housing market functioning as an investment market has an incomparable efficiency.

## **1.2 Regulation and Stimulation Policy in the history**

In order to restrain the rapid rise in the housing price, the Chinese government has implemented several rounds of real estate regulation in the past two decades. (Zhao, 2016) The regulation can be roughly divided into two types: housing demand restraint and land supply adjustment. Under these two main approaches, there are additional methods for the government to use. To control the housing demands, the government can adjust interest rate and home down payments, implement purchase limitations, or provide different housing options like affordable housing.

To control land supply, the land use plan has been seen as a tool for the government to control real estate development. (Yang & Chen, 2014) The land use plan legislates the site selection proposition statement and the constructing-land development license. The land use plan is embedded in three of the major forms of planning procedure: Master Plan, Regulatory Plan, and Constructive-Detailed Plan. The government has the right to approve and advise the revision of the content of land use planning of sites, which mainly involves FAR (Floor-Area Ratio), architectural density, greening rate and road systems.

In the past few decades, the Chinese government has gone through several rounds of back and forth in regulatory directions. Below is a policy summary of the real estate control from 1997 to 2016.

Year	Control Policy	Year	Control Policy
1997	Implementation of moderately tight fiscal and monetary policy.	2007	Liquidation of land value-added tax; Raise of the down payment of second home to 40%; Strengthening of land supply and shortening of land development cycle.
1998	Promotion of real estate industry; Focus on developing affordable housing; Implementation of monetization of housing distribution.	2008	Combat land storage in development; Further raise of the deposit reserve ratio; Exemption tax on affordable housing; Reduction of threshold for real home buyers.
1999	Implementation of active fiscal policy; Encourage individual to redeem houses; Elimination of personal income tax; Start real estate market; Halve deed tax.	2009	Discount mortgage rate for credit loaner; Regulation of violation in planning FAR; Proposition of property tax; Supervision of idle land of real estate.
2000	Promotion of housing consumption; Exemption tax of housing accumulation funds; Reduction of lease revenue tax.	2010	Stop mortgage of the third home; Household purchase limitation; Stop real estate company from issuing stock and bonds.
2001	Promotion of digestion of overstocked housing; Increased investment in real estate;	2011	Promotion of smooth, healthy estate market; Increased the down payment to 60% for the second home; Enforced 70% land supply on affordable housing; Increased tax on recently purchased home.
2002	Prevention of commercial banks from loaning to designated insurance units; Reduction of housing loan interests rate; Strengthening of the macro-control of real estate market; Recovery of the land value-added tax; Strict control of the total land supply.	2012	Forbidding of villa in real estate market; Protection of rigid needs in housing, support first home buyers; Strengthening of supervision in housing funds; Strike of housing on ambiguous land property
2003	Strengthening of control over real estate mortgage; Increase of the down payment to 40% for the second home; Real estate tax on sales of the house.	2013	Strict curbing of speculative investment; Increased supply of ordinary residential buildings; Acceleration of affordable housing projects; Increased tax on second-hand housing transactions

2004	Combatting of investment behavior in Shanghai; Suppression of housing bubbles; Strict control of the land market.	2014	Increased financial support for affordable housing; Support of reasonable mortgage demand; Support of reasonable financial need of real estate enterprises; Relaxation of provident fund loans; Decreased mortgage interest rate.
2005	Cancellation of preferential mortgage; Real estate tax over house trading; Raise of regulation to the political level; Tightening of real estate trust.	2015	Exemption of business tax on transferring housing purchased more than two years; Reduction of down payment to 40% for the second home, 20% for the first home; Reduction of mortgage interest rate.
2006	Raising of loan interest rate; Restriction of supply type of housing; Regulation of foreign investment; Further regularization of land market.	2016	In cities without purchase restriction, reduction of down payment to 30% for the second home; Strict purchase limitation in 1 <sup>st</sup> tier cities; Differentiated housing credit policy.

**Table 1.2.1 Real Estate Control Policy (1997 ~ 2016)**

### **1.3 Social Media Penetration**

According to the forty-first “Statistical Report on the Development of China’s Internet Network” released by the China Internet Network Information Center, as of December 2017, the number of Internet users in China reached 772 million and the penetration rate reached 55.8%. (Zhao, & Shen, (Eds.), 2018). The number of mobile internet users in China reached 735 million, which takes up 97.5% of total usage. Under the most common circumstances, common users are the major forces in the social media content. Meanwhile, public organizations including policy-makers are also leaning to make influence through the internet. In 2017, the number of online government service users in China reached 485 million, accounting for 62.9% of the total internet users. With the penetration of internet, more and more citizens have become more likely to educate themselves about public affairs and public policy through internet. At the same time, people are also more likely to discuss public affairs through social media than before. In the foreseeable future,

the penetration of internet will keep growing. Social media provides policy makers with the opportunities to gauge public opinion with regard to certain policy issues. (Goerge, Ozik & Collier, 2015)

## **2 Research Objective**

The main objective of this research is to understand the relationship among online discussion, the real estate market, and controlling policy. To accomplish that, it is first necessary to show that one can use a robust method to retrieve and extract information from the internet and process them into an adequate format. One then needs to show a scientific approach to organize and structure these online discussion data for analysis. With the obtained data and some appropriate analysis, one might be able to answer the question regarding the interaction among public opinion, market situation, and policy direction in a statistical sense. Such as: What do people discuss about the market? How does market discussion shift with respect to time? Do people talk about different things in different market situations? How do different regulatory policies influence the market discussion? Do different market discussions influence each other?

The argument for this research is that online discussion is a useful method in designing market related policy. Based on the result of this study, policy-makers will have a powerful tool to gauge public opinions and market expectation in this modern age. This research is important because it will explore an innovative method on a data sources to complement traditional study in policy research. The difficulty of this research lies in the complexity of analyzing large scale, online unstructured data which have not been exploited in most studies of regional science and planning. Compared to regular market statistical analysis in market analysis, this research is



intended to find a reliable measurement that could benefit the social aspect of economics analysis in behavioral studies.

### **3 Concepts and Tools of Text Mining**

#### **3.1 Text Preprocessing and Representation:**

One of the most significant steps in analyzing unstructured text data is transforming it into a structured representation that is usable for statistical analysis. The most used method is ‘bag of words’ representation which involves converting the set of documents, which is usually referred to as a ‘corpus’, into a document-term matrix or a so called document-token matrix. The matrix is composed of rows representing documents and entries representing words in that document by sequence; or more commonly, just the words in that document along with number of occurrence in the document. Since the text data in this research is scraped from internet, some steps of removing unrelated characters and HTML web tags are needed when cleaning the data.

In this research, an additional step of Chinese word segmentation, which is the task of segmenting Chinese sentences into word sequences, is used. Since a word is the basic sense-carrying language unit in Chinese, and Chinese is not a naturally space-segmented language, word segmentation is the first step before text conversion into a matrix representation. The software package used for this task is a cloud based package, Language Technology Platform Cloud (LTP-Cloud, <https://www.ltp-cloud.com/intro/en/>) developed by the Research Center for Social Computing and Information Retrieval at Harbin Institute of Technology. LTP's Chinese word segment module is based on a machine learning framework that is powerful enough to solve the ambiguity problem in segmentation tasks.

### 3.2 Topic Modeling:

Topic modeling is a machine learning and natural language processing algorithm that can automatically summarize large collections of text by discovering the latent ‘topics’ found within that set of documents. (Han, Mankad, Gavirneni & Verma, 2016) Latent Dirichlet Allocation (LDA) is a widely used topic modeling method in this study. It is a generative model which finds topics with probabilistic frameworks. (Blei, Ng & Jordan, 2003). The idea behind this approach is that all documents share a set of topics and each topic is composed by a collection of words that are more likely to be used in the discussion of this topic. Each document, nonetheless, has a different probabilistic mixture of those topics in the shared set. Meanwhile, each topic is also a probabilistic mixture of different words used in the entire corpus of documents. In another sense, certain words are more likely to be used in a certain topic. The results of the LDA topic modeling are composed of two parts:  $P(\text{topic}|\text{document})$  which is the probability distribution of topics given a certain document, and  $P(\text{word}|\text{topic})$  which is the probability distribution of words given a certain topic.

The software package used in this research is MALLET (MACHINE Learning for Language Toolkit. <http://mallet.cs.umass.edu/index.php>) created by researchers at the University of Massachusetts Amherst, which is a Java-based package for statistical natural language processing, document classification, clustering, topic modeling and information extraction. The software can automatically read in proper segmented documents, transfer them into matrix representation and train on the topic. The most important input for the topic modeling task is the number of resulting topics, which will determine how many topics will be generated in the training result. The software can also screen out stop words from documents. A stop word is a commonly used word, such as ‘the’, that can be ignored in the task. The intuition behind screening stop words in topic modeling

is to remove the words that are commonly used, with little contribution in understanding the topic, so that we can better interpret the meaning behind a topic with words containing more information.

### **3.3 KeyGraph:**

KeyGraph is a graph analytical approach for fast topic relation detection used in this research. The software used in this study is Polaris which is developed by Ohsawa Laboratory on Chance Discovery at The University of Tokyo. (<http://www.panda.sys.t.u-tokyo.ac.jp/KeyGraph/>) KeyGraph was initially proposed by Yukio Ohsawa along with the chance discovery theory. (Ohsawa, 2003). KeyGraph is a data visualization tool for creative system design that can analyze and synthesize data and human's ideas on consumer's behaviors, conversation, earthquakes and medical treatment etc. The purpose of Ohsawa in developing this tool was to make data and thoughts visible to support the theory of chance events which rarely happen but are linked to a daily event. The chance theory is very helpful in unveiling uncertain risks or opportunities. In this study, however, the chance theory will not be used because the study objective is to find the relationship between market and topics discussed on social media. The chance item mainly refers to a rare but significant topic among different topics which does not align with the interest of this research. Furthermore, it is even unreliable to test chance theory on this data since we do not want to make further inferences until we can ensure the discussion on the internet is related to situations in the real estate market, which is the focus of this study.

Even so, the KeyGraph tool is still powerful in terms of finding the relationship between topics and clustering the topics. To construct a KeyGraph to visualize topic relations among topics, high frequency topics are first extracted, which are generated from top N topic from a sorted frequency list of the whole topics set. These selected topics will be represented by a black node in the graph. Next, it is necessary to measure the relationship between topics. In topic detection

studies (Ko, Jeong, Choi & Yoon 2018), the Jaccard coefficient is commonly used to represent the co-occurrence frequency of two topics. These topics-pair relationships are then sorted by the Jaccard coefficient. The top M relationships will be formed in between these topic-pairs. In the graph it is represented by black edges connecting topic nodes. By doing so, a relationship graph is outputted that could help identify the topic clusters. The number M and N are the input for the software which can be assigned to suit the research purpose.

The formula for Jaccard coefficient is:  $J(T_i, T_j) = \frac{\text{Freq}(T_i \cap T_j)}{\text{Freq}(T_i \cup T_j)}$ . In this particular study,  $\text{Freq}(T_i \cap T_j)$  is the frequency of topic  $T_i$  and topic  $T_j$  co-occurring in the same document, and  $\text{Freq}(T_i \cup T_j)$  is the frequency of topic  $T_i$  or topic  $T_j$  occurring in documents.

## 4 Data Preparation

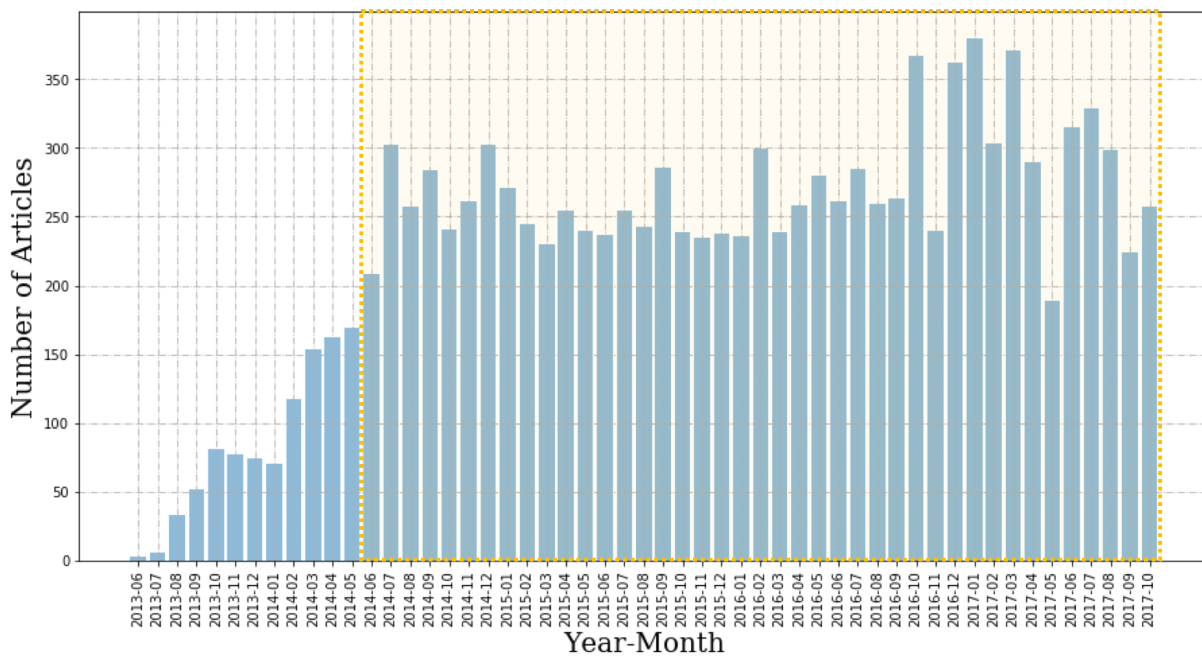
### 4.1 Document Source and Data Wrangling:

The WeChat Subscription Account platform is an affiliate platform attached to WeChat, one of the largest standalone messaging apps as of 2018. It was developed in Aug 2012. The reason for choosing this platform as the data source is that it is currently the most widespread social media platform for any entity to register and promote ideas. As described in the official website: ‘WeChat Subscription account is typically the most basic choice of the official accounts. It allows you to push frequent content to your followers. Account manager can broadcast one message per day. The account followers will see the update information in the subscription area.’ Unlike other media applications, WeChat Subscription Account platform provides direct access to articles to be directly forwarded and shared to the group chat and the WeChat moment for WeChat users to see. Links from other platforms need to be viewed in a WeChat app which creates a barrier for users

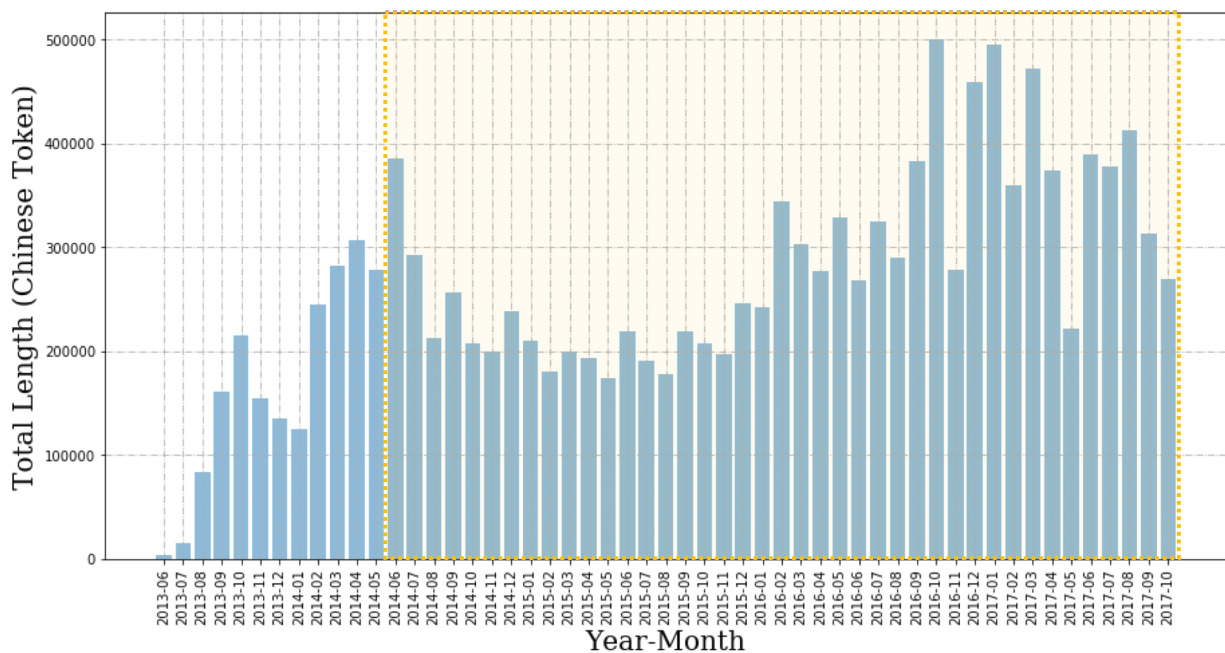
to browse outside of WeChat. However, it does not prevent WeChat from being the most popular social media platform in China. According to Tencent Penguin Smart, the total number of global active users has reached 889 million. 45% of the users have more than 200 WeChat friends. Among over 10 million WeChat Subscription Accounts, nearly 30% of them have more than ten thousand subscribers.

The document data are web-scraped from the WeChat Subscription Account platform and each article is saved as a separate document. The metadata fields for a document include title, author, date, and content. Since WeChat articles normally can be only viewed in App, the Sogou search engine was also used for WeChat platform to search for the Wechat articles. The key word used in the article search engine is “房价” (Housing price). For each month from June 2013 to October 2017, the top K numbers of articles from the search result were scraped as the sample in order to estimate the real online discussion. K varies from month to month because some months in 2013 have relative low total amount of articles and the Sogou search engine has a random verification process which can interrupt the scraping process and prevent me from getting the identical amount of articles for each month. The average number of articles per month is around 270. When K grows to over 360, the scraping program was manually stopped since the search engine sometimes failed to provide relative articles from this point because of algorithm instability. All the articles were kept as more data can contribute to a more robust topic training result. Since the variables will be averaged by the number of articles in that month, the final estimation of topic result will not be affected as long as the sample size is sufficiently big (If search engines are assumed to provide a random result, based on 99% of the confidence level, the population of an average of 1000 articles, a sample size of 143 is needed to get an estimate with a confidence interval of 10). Duplicated articles in the same month were removed. No data were scraped before

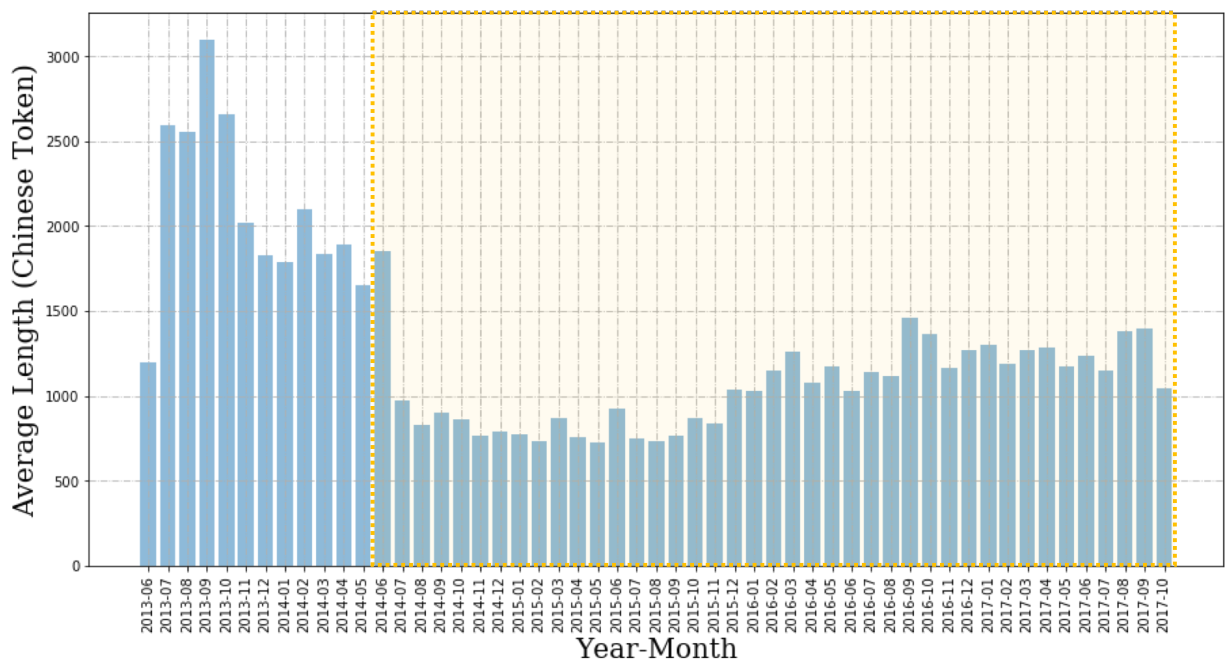
June 2013 because of an insufficient amount of articles. The figures below show the number of articles scraped per month and the average length of article per month. Although all the documents were included into the topic modeling training process, only months from June 2014 to October 2017 were selected for statistical analysis as previous months cannot provide the sufficient number of articles necessary for statistical inference.



**Figure 4.1.1 Number of Articles Scraped per Month**



**Figure 4.1.2 Total Length of Articles per Month (in Chinese Token)**



**Figure 4.1.3 Average Length of Articles per Month (in Chinese token)**

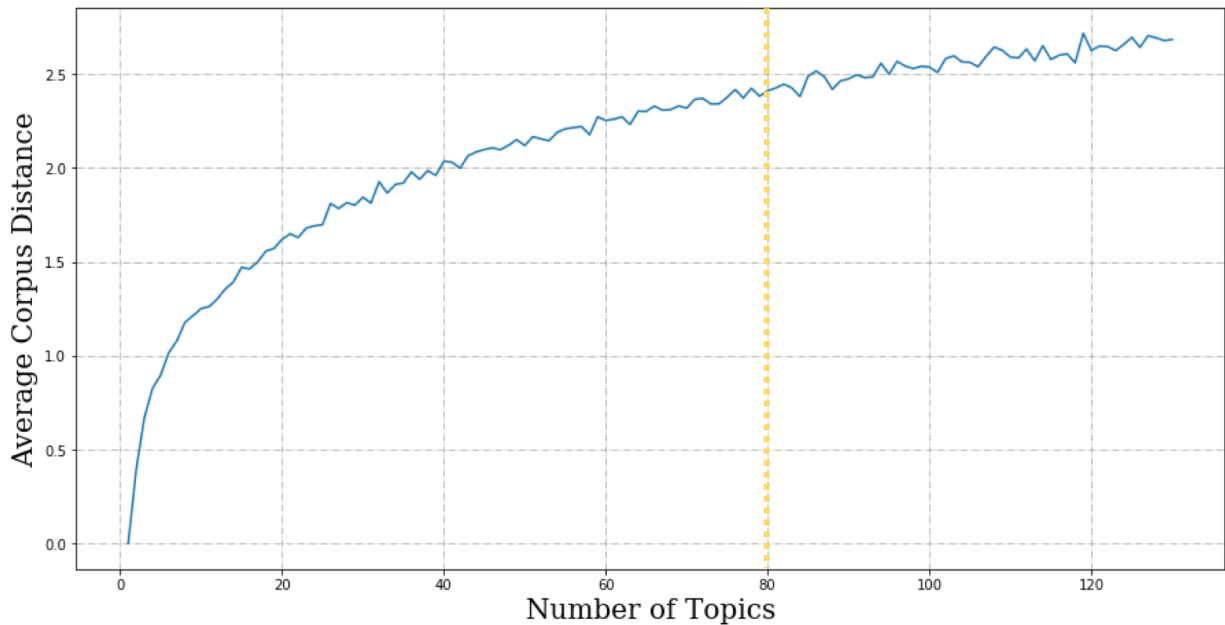
## 4.2 Topic Modeling:

To prepare data for the topic modeling package, word segmentation and unrelated character removal were implemented for each of the document. All the special characters, English characters, and numbers were removed during the tokenization process after which each word token was considered as a unit for matrix representation of documents. In MALLET topic modeling options, the sequence was chosen to be kept within the data. Since one-character or two-character words are common in Chinese, the token-regex was set to "`\p{L}+`" that allows words with any number of characters. The Chinese stop words used combine a set of commonly used words and some additional words that do not contribute to the housing market discussion semantically.

In determining the appropriate number of topics, several methods are useful including the elbow method, perplexity score, information criterion method and information theoretic method. The elbow method was chosen for this study as it is a method of interpretation and validation of consistency within cluster analysis. The main intuition behind the elbow method is to find the turning point where the marginal gain of the measurement metric will drop, giving an angle in the graph. The metric chosen was the ‘corpus distance’ generated by MALLET diagnostics. Corpus distance measures how far a topic is from the overall distribution of words in a corpus, which is the result of the model trained with the one topic setting. Since choosing a proper number for topic amount is a tradeoff between topic distinction and topic interpretability, there was a need to find an elbow point to balance both sides of effects. The topic model was thus trained with a number of topics from 1 to 130 and plot their average corpus distance in the figure below. There is no idealistic turning point in this graph since the corpus distance smoothly converges to a certain scale. The reason for the shape of such a smoothly growing graph might be due to the corpus being relatively huge and miscellaneous. Finally, the number 80 was chosen as the input for topic



modeling as there is an observable increasing fluctuation after 80, representing the meaning behind topics becoming ambiguous and topics with similar meaning starting to appear more often.



**Figure 4.2.1 Elbow Method for Determining Appropriate Amount of Topics**

The result of topic modeling with a topic number of 80 includes a probability distribution matrix of topics given documents, and a probability distribution of words given topics. An example of words given topics matrix are shown below. The matrix of topics given documents will be exhibited later, filtered with a probability threshold.

Topic ID	Topic Name	Words (Probability Top 10)
0	Company Issues	Company(0.02314), Yuan(0.01844), Wenzhou(0.01399), Enterprise(0.01282), Corporate(0.01255), Developer(0.001003), Money(0.00789), Funding(0.00758), Contract(0.00669), Home(0.00665)

1	Construction	Area(0.03813), Building(0.02940), Square meters(0.02291), Floor(0.01564), House plan(0.01259), Engineering(0.0113), Flat meter(0.01085), Residential(0.00966), Total(0.00825), High(0.00824)
2	China Development	China(0.03386), Economy(0.01472), Development(0.01329), Social(0.01146), Future(0.01073), Becoming(0.01046), Corporate(0.00804), Problem(0.00691), Country(0.00684), Wealth(0.00639)
3	Pearl 1 <sup>st</sup> *	Shenzhen(0.19833), Housing price(0.02699), Wan*(0.01856), Dongguan(0.01552), Shenzhen city(0.01023), Longhua(0.00975), Huizhou(0.00958), Yuan(0.00909), District(0.00758), Deep(0.00729)
4	Pearl not-1st	Yuan(0.05603), Average price(0.03316), Room(0.02740), Address(0.02186), Zhuhai(0.01709), Boulevard(0.01612), Discount(0.01465), Sale(0.01303), Property(0.01220), Intersection(0.01213)
5	Price Rank	House(0.0516), Price ranking(0.01963), Cities(0.01507), Rank(0.01448), City(0.01228), Nation(0.01195), No.(0.01156), Average price(0.01118), China(0.01042), County(0.01035)
6	House Agents	Intermediary(0.02813), Second-hand housing(0.02775), Price(0.02116), Million(0.01929), Customer(0.0182), Owner(0.0181), House(0.01618), Residential(0.01569), Sale(0.01520), Room(0.01403)
7	Company News	Real estate(0.03701), Real estate(0.02113), Company(0.01777), Corporation(0.01515), Enterprise(0.01507), Project(0.01455), Industry(0.01434), Vanke(0.01121), Group(0.0099), Yuan(0.0096)
8	Price Tease	House(0.02878), Buy(0.02655), Million(0.02215), Money(0.01425), Sell(0.01108), Dollar(0.01088), Set(0.00927), RMB(0.00878), Price(0.00841), Approx(0.00829)
9	Price Trends	House Prices(0.07657), Down falling(0.02353), Falling(0.02087), Property Market(0.01255), Uprising(0.01036), Market(0.00999), Price Drop(0.00955), Rising(0.00945), Appearing(0.00943), Already(0.00863)

Notes: ‘Pearl 1<sup>st</sup>’ = Pearl delta area 1<sup>st</sup> tier cities, ‘Wan’ = ten thousands

**Table 4.2.1 Example of Topics and Words with High Probability**

In order to effectively analyze and identify each topic, a name was assigned for each of the topics. The naming process is conducted through human interpretation based on the top words with high probabilities in the topic-words distribution matrix. There is an unavoidable subjectivity when transferring machine interpretation into human interpretation. The meaning behind names will also be used in the later topic clustering. The list of 80 topics and their names are shown in the table below.

ID	Topic Name	ID	Topic Name	ID	Topic Name	ID	Topic Name
0	Company Issues	20	Service Industry	40	Specialist Analysis	60	Trade Log
1	Construction	21	Negative Analysis	41	Listing-2	61	Cities Names
2	China Development	22	National Price	42	Policy Announcement	62	Career Plan
3	Pearl 1st	23	Mechanics Analysis	43	Money Supply	63	Interior Design
4	Pearl not-1st	24	Control Policy	44	North-China Cities	64	School District
5	Prince Rank	25	Life Plan	45	Oversea Market	65	Professional Economic Analysis
6	House Agents	26	Anecdote	46	Poor-Story-1	66	Immigration
7	Company News	27	Procedure Supervise	47	Land Finance	67	Financial Analysis
8	Price Tease	28	Headline Trash-News	48	Economic Analysis	68	Capital-Area Cities
9	Price Trends	29	Market-Analysis-2	49	Land Bidding	69	Oversea Bubbles
10	Travel Island	30	Contract Dispute	50	East-China Cities	70	Oversea Properties
11	Price Complaint	31	City Culture	51	Urbanization	71	Account Promotion
12	Market-Analysis-1	32	Other Price	52	Poor-Story-2	72	Life Story

13	Mortgage	33	Tier Price	53	Areal Service	73	Market-Analysis-3
14	Used Price	34	SE-Coastal Cities	54	Listing-3	74	Purchase News
15	Listing-1	35	South-west Cities	55	Shanghai	75	New-District Plan
16	not-1st Cities	36	Vulgar Story-1	56	Middle-Class Issues	76	Purchase Limitation
17	Positive Analysis	37	Salary Expense	57	South Yangtze	77	Vulgar Story-2
18	Anhui	38	Hong Kong	58	Tier-Market Analysis	78	Property Tax
19	Increasing	39	Travel Domestic	59	Bride-price Story	79	Investment Analysis

**Table 4.2.2 List of Trained Topics and Corresponding Names**

Originally in the topics given document matrix, each document has every topic with its contribution probability. It is usually the case, nonetheless, that one document (WeChat article) has a few topics. In order to determine an exact number of topics for each document for the input of KeyGraph analysis, the threshold on the topic distribution model was used to determine whether a topic has considerable probability to be contained in the discussion of that document. Essentially, one needs to find a proper threshold  $\alpha$ , and assign a certain topic to a certain document to that  $P(\text{Topic}|\text{Document}) > \alpha$ . After we obtain a threshold, we can assign topics to documents. An example of assignment is shown in the figure below. The methods for choosing proper threshold value will be discussed in next sector.

Document ID	Topics (with Threshold $\alpha = 0.18$ )
2013-07-1	Career Plan (0.885206422018)
2013-07-2	Price Tease (0.193882042254)
2013-07-3	Life Story (0.982379415761)
2013-07-4	Immigration (0.796669947507)
2013-07-5	Career Plan (0.28125), Life Story (0.219119822485)
2013-07-6	Construction (0.761677046263)
2013-08-1	Company Issues (0.426358628373)
2013-08-2	School District (0.263981835564)
2013-08-3	Life Plan (0.295997191011), Life Story (0.294124531835)
2013-08-4	Construction (0.868125)

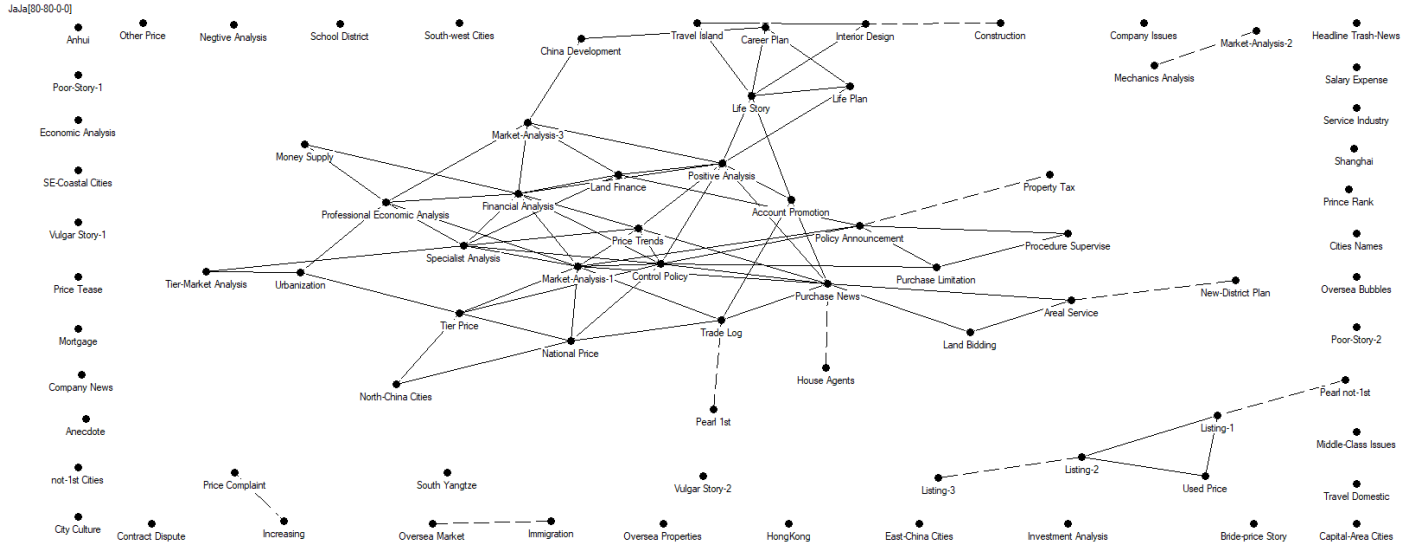
Notes: 'high probability' =  $P(\text{Topic}|\text{Document}) > \alpha$ , where  $\alpha = 0.18$

**Table 4.2.3 Example of Document and Topics with High Probability**

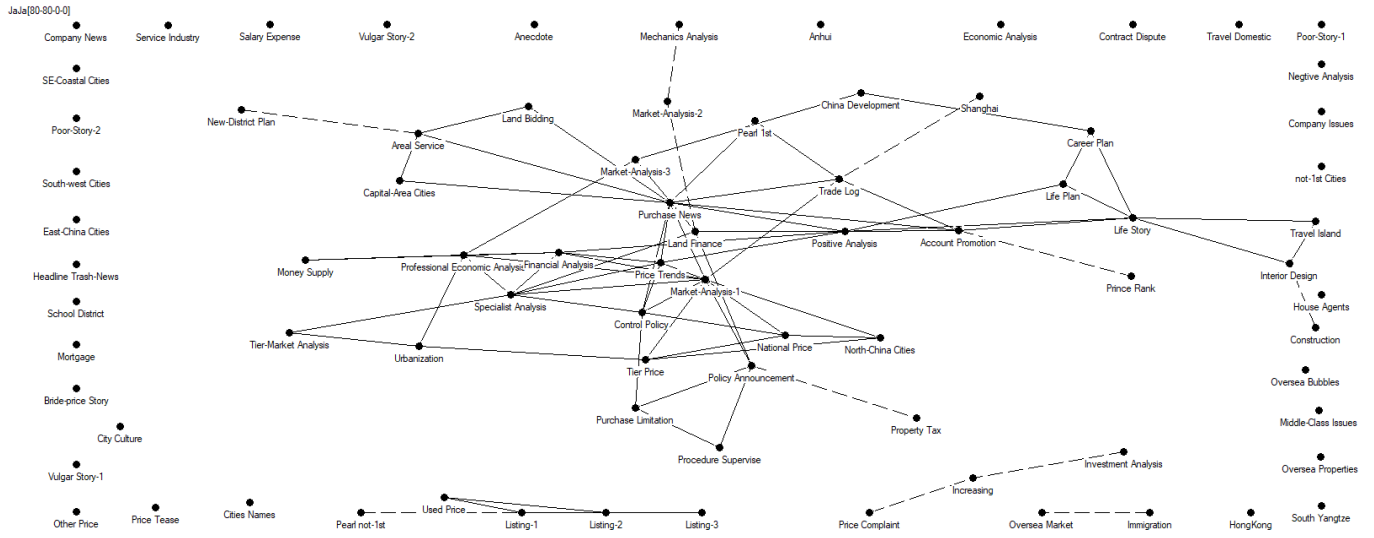
### 4.3 Construct KeyGraph and Clustering Topics:

The purpose for using KeyGraph in this study is to support topic clustering with statistical observation. The input for KeyGraph construction is the co-occurrence between pairs of topic in a single document (if a document has more than two topics assigned). The co-occurrence matrix can be obtained by setting a threshold probability to filter out the unlikely topics with low probabilities. For a distinct threshold value, therefore, a distinct topics co-occurrence matrix and corresponding KeyGraph can be generated. A fixed value of M (number of black nodes) and N (number of black edges) can be used that equals to the number of topics of 80 because every node needs to appear in the graph with an adequate amount of co-occurrence relationship. To find a proper threshold, threshold values of 0.06, 0.09, 0.12, 0.15, 0.18, 0.21 were tested to generate a corresponding KeyGraph and topic number statistics.

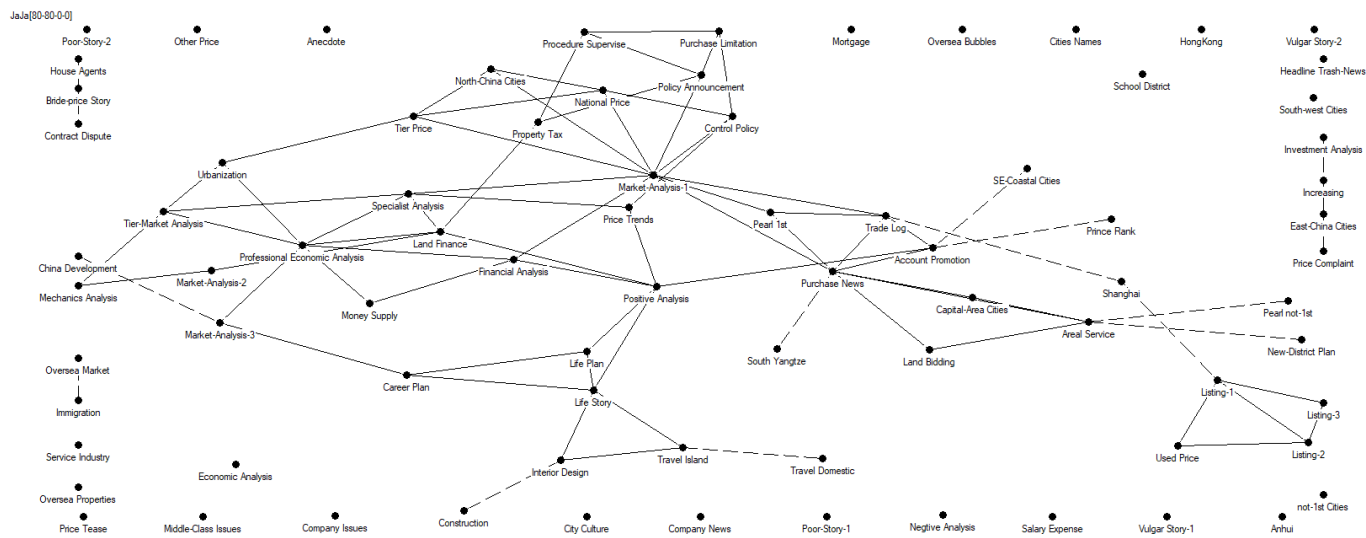
## Topic KeyGraph Generation, M=80, N=80



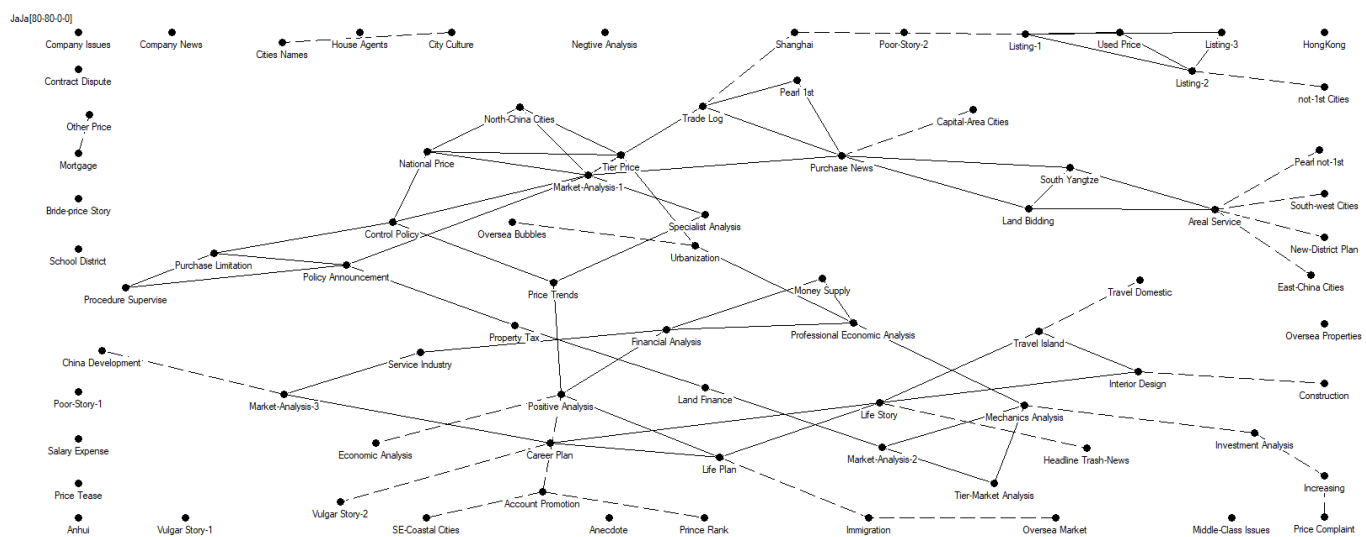
Topic Threshold  $\alpha = 0.06$



Topic Threshold  $\alpha = 0.09$



**Topic Threshold  $\alpha = 0.12$**



### Topic Threshold $\alpha = 0.15$





tendency explains why the cluster relationship is denser in the KeyGraph with a smaller threshold as topics are more likely to co-occur and vice versa. At the end, a threshold of  $\alpha = 0.18$  was chosen because the cluster in KeyGraph is relatively spread out and the number statistics seem reasonable considering the content length of articles.

Threshold	Min # of T/D	Max # of T/D	Average # of T/D	Zero Topic Document Percentage
$\alpha = 0.06$	1	9	4.25	0 %
$\alpha = 0.09$	0	7	3.03	0.016%
$\alpha = 0.12$	0	6	2.27	0.313%
$\alpha = 0.15$	0	5	1.77	2.284%
$\alpha = 0.18$	0	4	1.40	7.387%
$\alpha = 0.21$	0	4	1.14	14.089%

Notes: ‘#’ = number, ‘T/D’ = Topics / Document

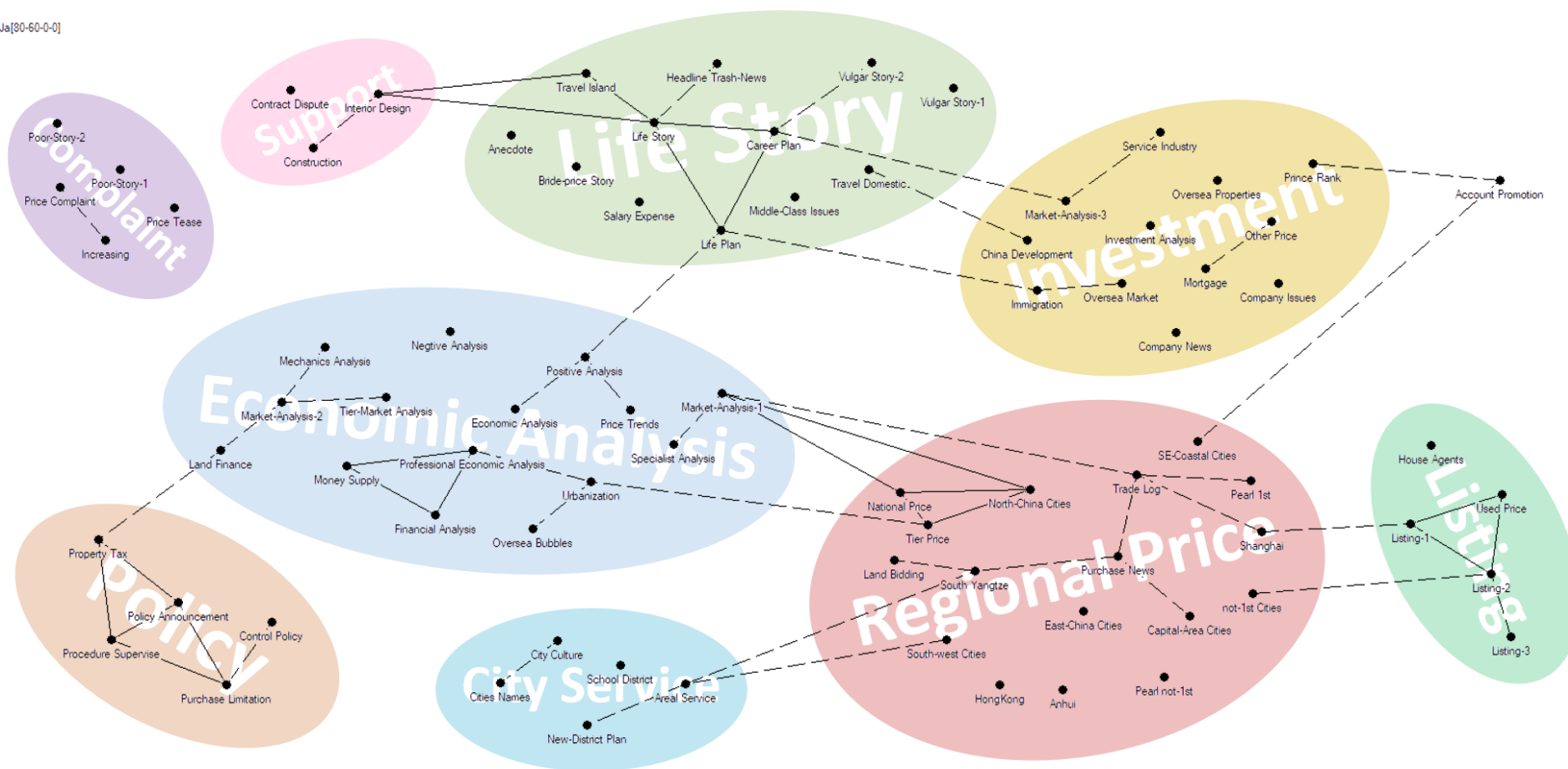
**Table 4.3.2 Different Thresholds and Corresponding Topic Properties**

While KeyGraph measures the likelihood of two topics being discussed together in a single document, it should not be the only rationale behind clustering topics. The co-occurrence results from both content similarity and reader similarity. Since the WeChat articles are published by different subscription accounts, two topics’ co-occurrence could be either because their contents are correlated or because they are designed to be seen by a reader group with similar interests (similar reading focus or similar region focus). In this research, content similarity is more important. Meanwhile, some topics are not linked in the KeyGraph but are related to a cluster in the content sense. In order to balance these effects, the KeyGraph results were manually scrutinized and clusters were built based on human understanding, but without largely revising the KeyGraph relationships.

The clustering results contain nine clusters (which will be referred to as the main topics for the rest of this paper). The contribution probabilities of the nine main topics add up to around 90%~95% in each month instead of up to 100% because the ‘Account Promotion’ topic was left out, which contributes to around 5%~10% of the topic each month. The naming rationale and the clustering graph are shown below.

Main Topic Name	Explanation
Complaint	Topics that contain negative, even cynical opinions towards the housing price problem.
Support	The support industries of real estate including interior design, construction, and stakeholder’s dispute.
Life Story	Topics that related to life plan, career, vacation, or stories involving housing price.
Investment	Topics related to company, market, price. Unlike Economic Analysis, this topic focuses on the view of the private sector or individuals.
Economic Analysis	Topics involving the analysis of the macro situation of the real estate market mostly focusing on formal economic analysis.
Regional Price	Topics related to price report and update to a certain area in China.
Listing	Most are pure properties information including property name, location and price. The properties could be first hand or second hand.
Policy	Topics involved with tax, regulations and policy discussion.
City Service	Topics that relate to service ability of regions or comparison of cities.

**Table 4.3.3 Naming Explanation of Clustered Topics (Main Topic)**



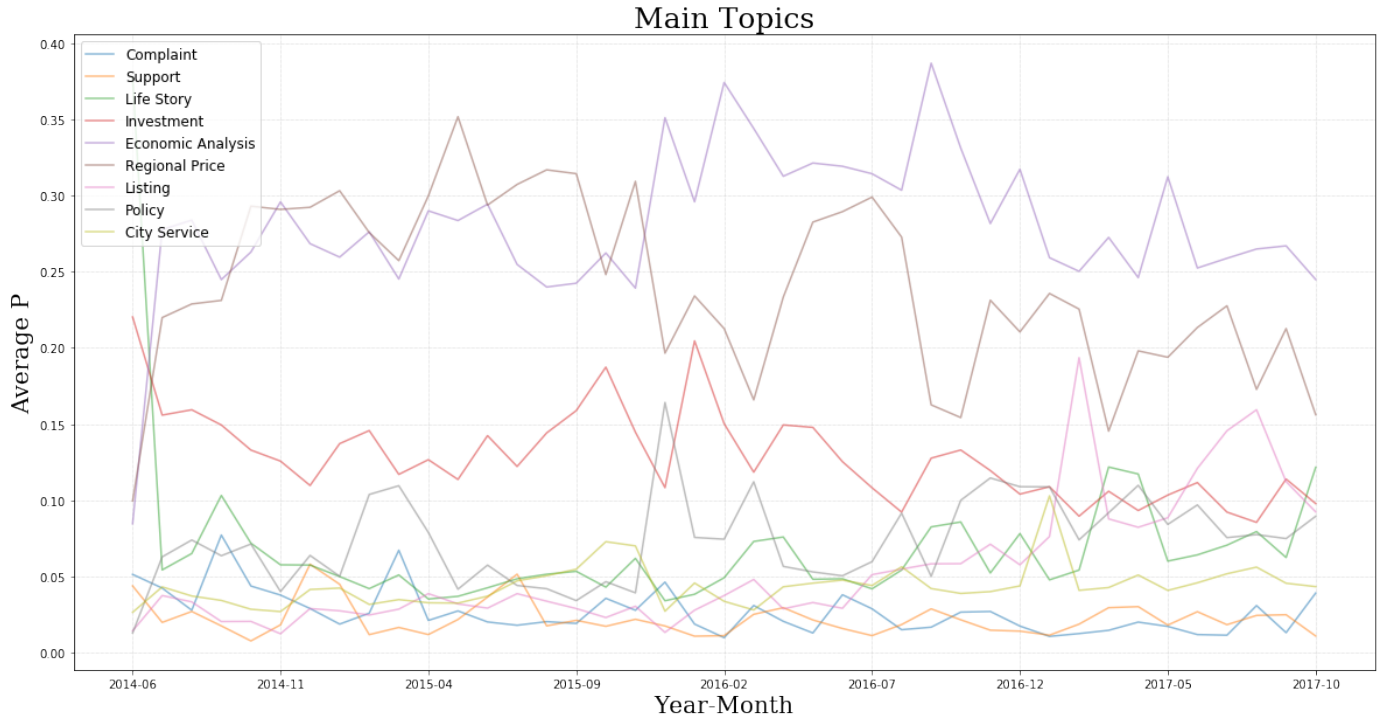
**Figure 4.3.1 Clustering KeyGraph with Nine Main Topics**

## **5 Regression Analysis - Static Model**

### **5.1 Summary Statistics on Topics and Main Topics**

After clustering, these nine main topics were used as the research subject for the statistical analysis. In measuring the extent of each main topic, the sum of the contribution probability was used instead of topic counts in the co-occurrence matrix. Even two documents with the same topic assigned can have very different contribution probabilities as long as they are over the threshold. The exact contribution probability is thus a better measurement of how much a certain topic has been discussed in the collection. Every single topic in a certain month was normalized by dividing the sum of contribution probabilities by the number of documents in that month.

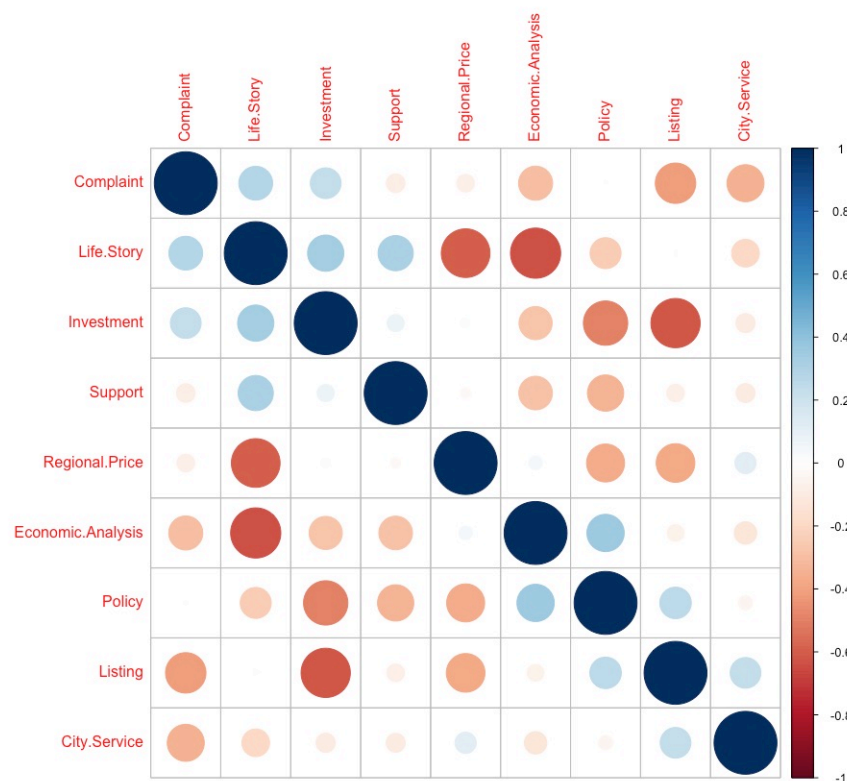
The line graph of the contribution probabilities of main topics and nine sub-topics are shown below, as is the correlation matrix of nine main topics. The time range of analysis is from June 2014 to October 2017 (41 months). In the graph of sub-topics, one can observe that there is an amazing collinearity among the sub-topics under a certain main topic. It seems that KeyGraph successfully captures this collinearity of contribution probability although its initial purpose was to measure the co-occurrence to which it might be related.



**Figure 5.1.1 Average Contribution Percentage of Main Topics per Month (2014-06~2017-10)**

In the correlation matrix of nine main topics, there are positive correlations between ‘Life Story’, ‘Complaint’ and ‘Investment’. This could be seen as an effect of unequal distribution in the real estate development where the difference between upper class and lower class grows larger and larger. This is a reflection on the internet discussion being heated between the two classes. There is also a positive correlation between ‘Policy’ and ‘Economic Analysis’, which is explainable in a sense that more analysis should be published when there is a policy change. There is also a positive correlation between ‘Listing’ and ‘Policy’, due to house listings being more likely to appear when general market expectations are positive, which will incur regulation policy. Although there could also be some policy discussion during the stimulus policy period, one might deduce that people are more likely to discuss policy issues during regulation than stimulation.

There are some significant negative correlations between ‘Life Story’ and ‘Regional Price’, and between ‘Life Story’ and ‘Economic Analysis’. This is because there are more topics in ‘Regional Price’ and ‘Economic Analysis’ when the real estate market is expanding. Topics about ‘Life Story’ are more likely to appear in a post-expansion time period. There are also some negative correlations between ‘Investment’ and ‘Economic Analysis’, ‘Policy’, and ‘Listing’. The reason behind this might be that people are more likely to invest or discuss business when a policy situation is stable and vice versa.



**Figure 5.1.2 Correlation of contribution percentage of main topics**

In the correlation graph one can observe that there are roughly two clusters of topics based on their similarities of correlation. In the upper half of the vertical coordinate, one can see topics of ‘Complaint’, ‘Life Story’, ‘Investment’, ‘Support’. These topics can be dubbed the subjective topics because they are all based on or influenced by personal feeling, choice, or opinions. In the

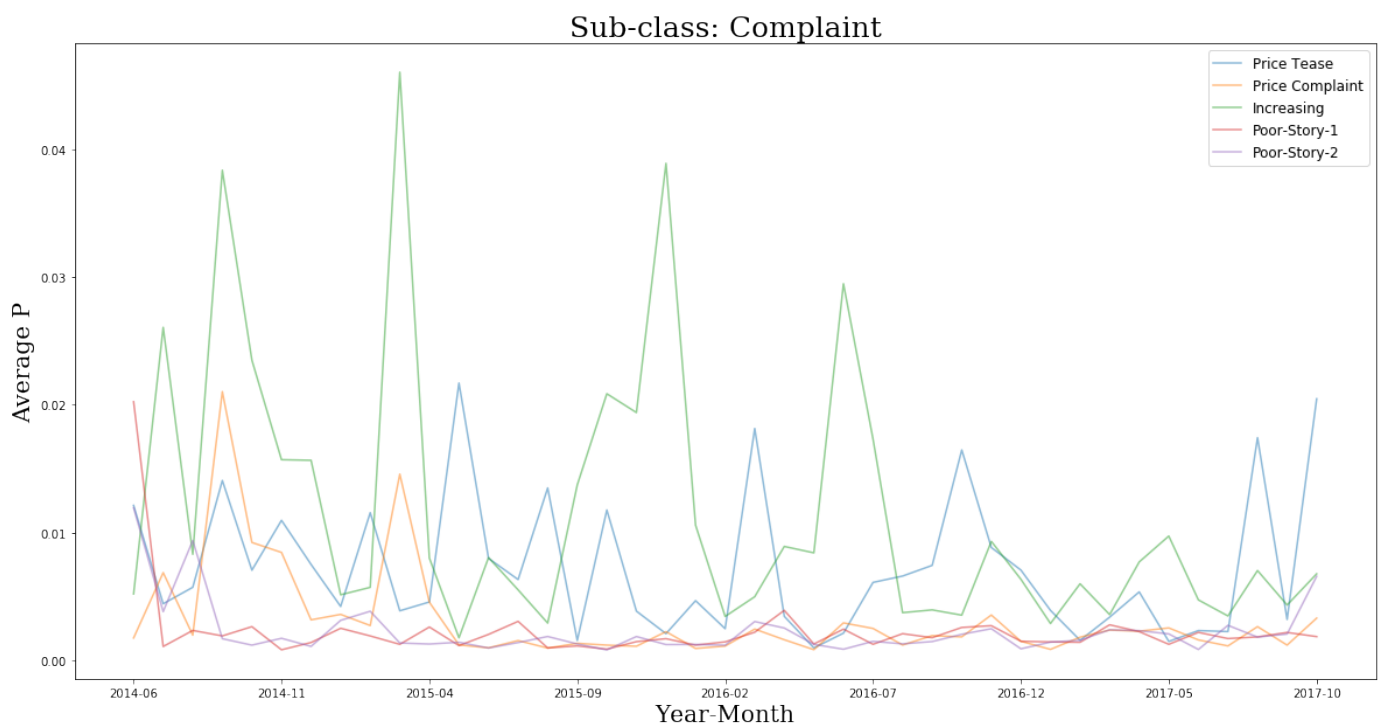
lower-half of the vertical coordinate, one can see topics of ‘Regional Price’, ‘Economic Analysis’, ‘Policy’, ‘Listing’, ‘City Service’. These topics are dubbed the objective topics because they focus on the objective sides of the real estate market that remain mostly uninfluenced by personal feelings or opinions. From the correlation plot one can realize the following: (1) the subjective topics have a mostly positive correlation with other subjective topics and a mostly negative correlation with the objective topics, and (2) the objective topics have a mostly positive correlation with other objective topics and have a mostly negative correlation with the subjective topics.

Statistically, if there are only two topics whose percentage add up to nearly 1, one can observe negative correlation because one is linearly correlated to another. If there are only two topics, they must be negatively correlated because the percentage of one topic is 1 minus the percentage of the other topic. The interesting finding is that the nine main topics have clustered into objective and subjective groups where each topic is roughly positively correlated with other topics in the same group and roughly negative correlated with other topics in the different group. From this observation, it might be suspected that the market condition might generally have a similar effect on the subjective topics and a correspondingly similar effect on the objective topics. It can be inferred that there might be a network effect on real estate buyers where most of the potential buyers have similar market expectation, which might be the reason a consistent distribution is observed within objective and subjective topics of market discussion.

## **5.2 Summary Statistics on Topics and sub Topics**

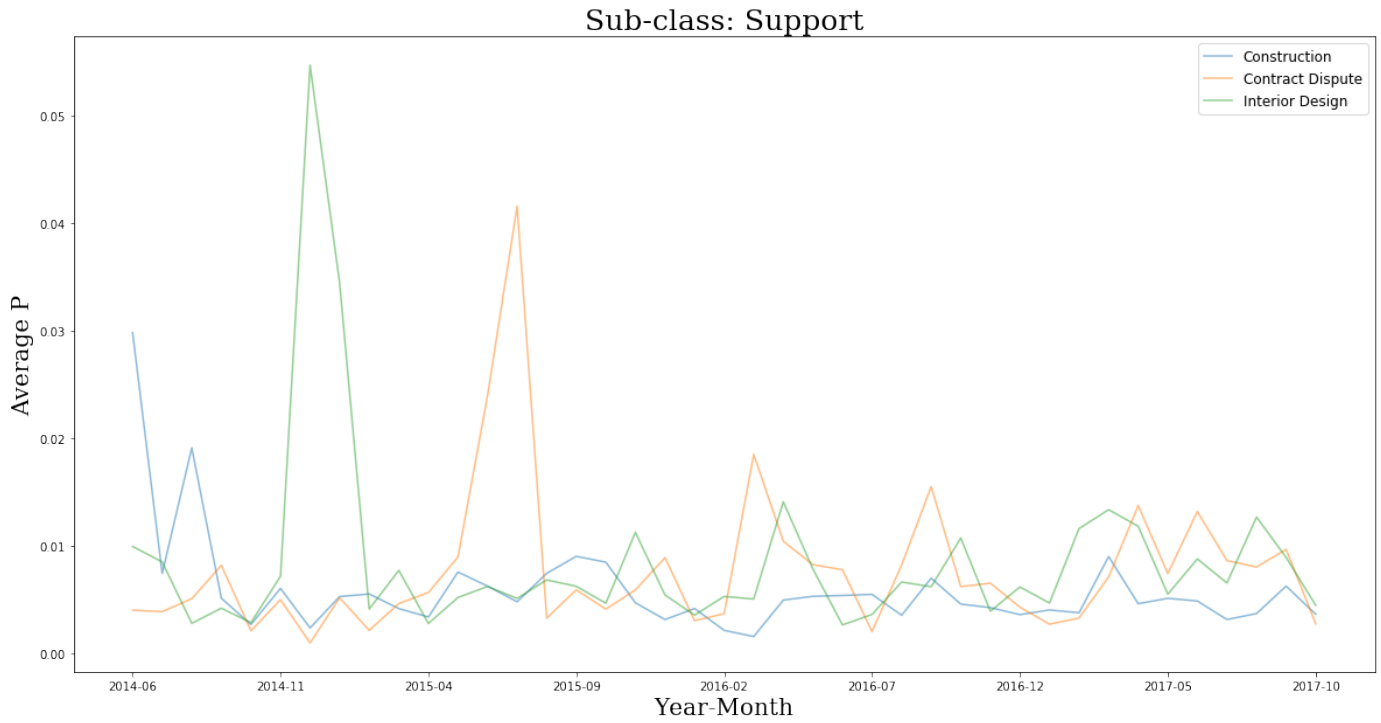
By taking a close look at each trend of the main topic or sub-topic, it is also possible to yield some very interesting findings that might be related to particular events in the past. From the proportion trend of main topic class, it is possible to observe a clear pattern of time series data

from which one can deduce the possible existence of a periodical pattern within these topics. The underlying reason might be its relationship to the market data which is also a time series, or it may be due to a recurring topic effect on social media. Around three major policy changes, one can observe some relatively big change in the proportion of topics. Instead of making a detailed observation, some regression analysis was performed to generally test if such a topic trend is statistically related to market condition to make a relatively reliable statistical inference using these graphs.

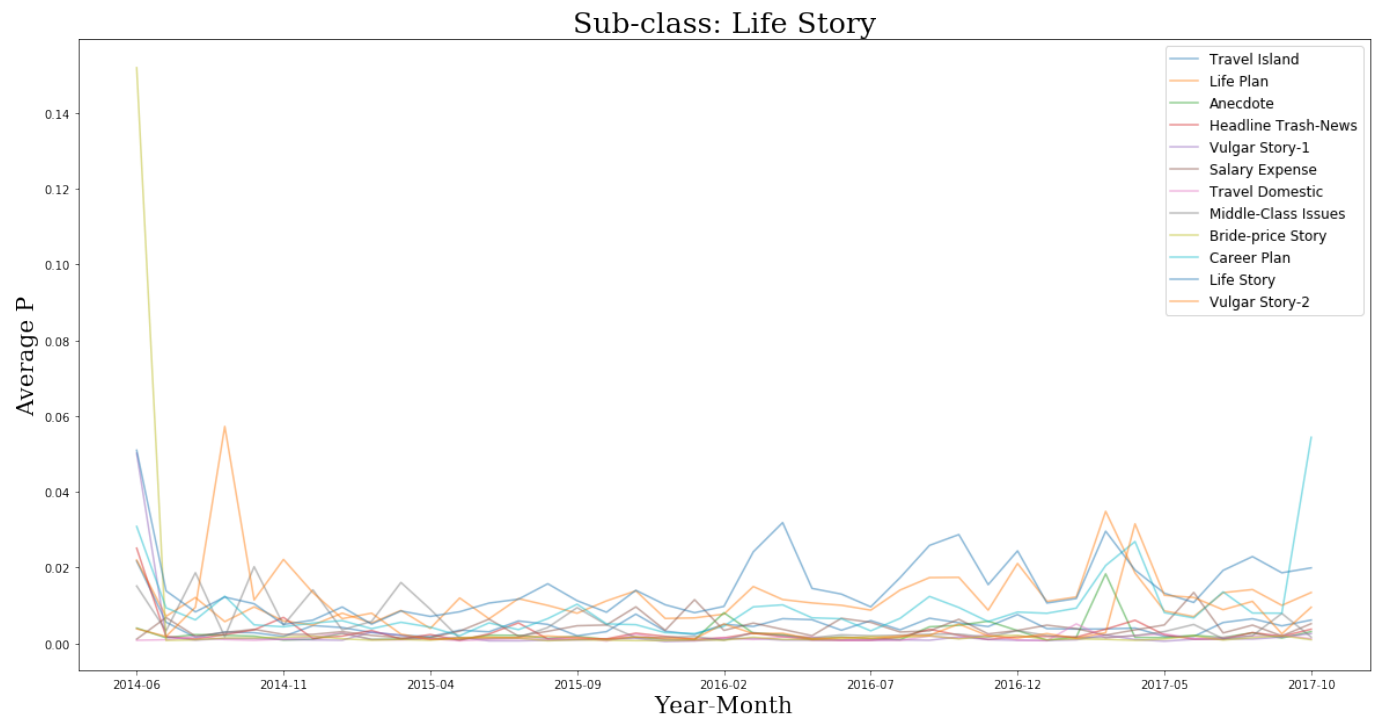


**Figure 5.2.1 Average Contribution Percentage of ‘Complaint’ Sub-Topics per Month**

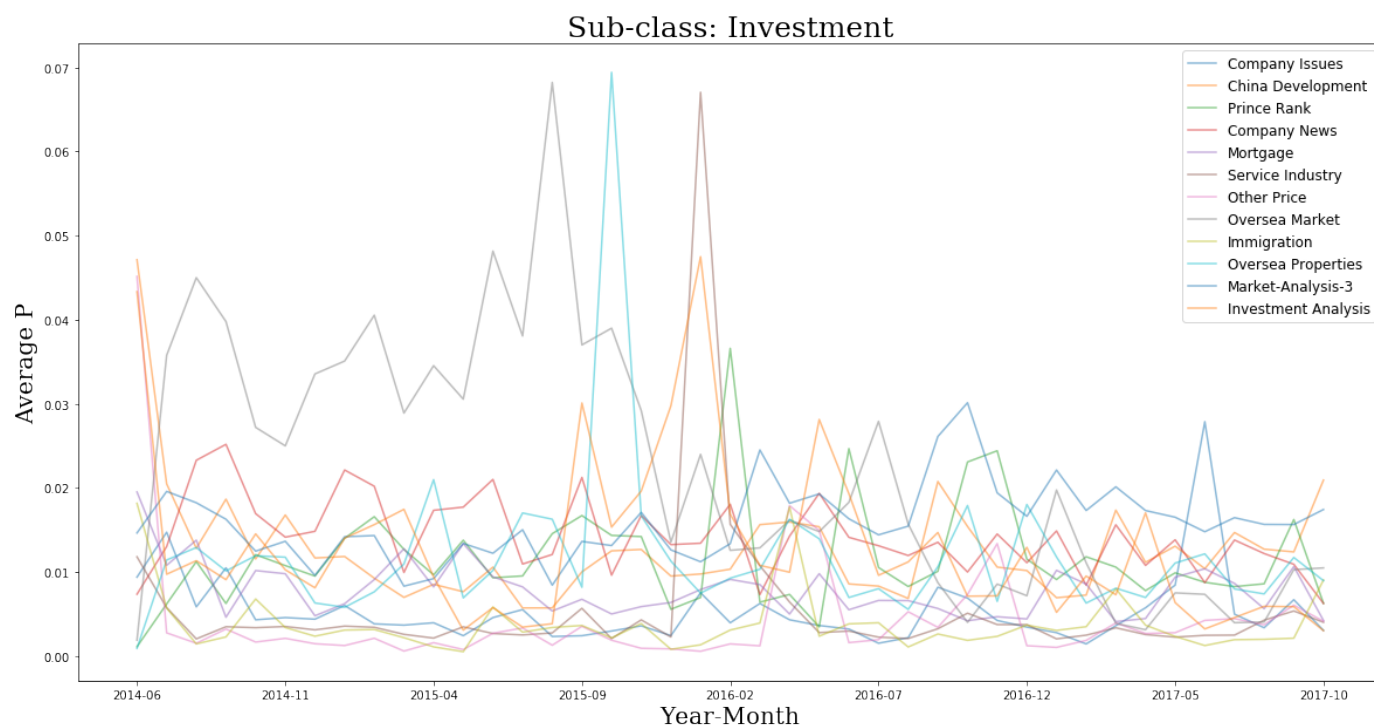




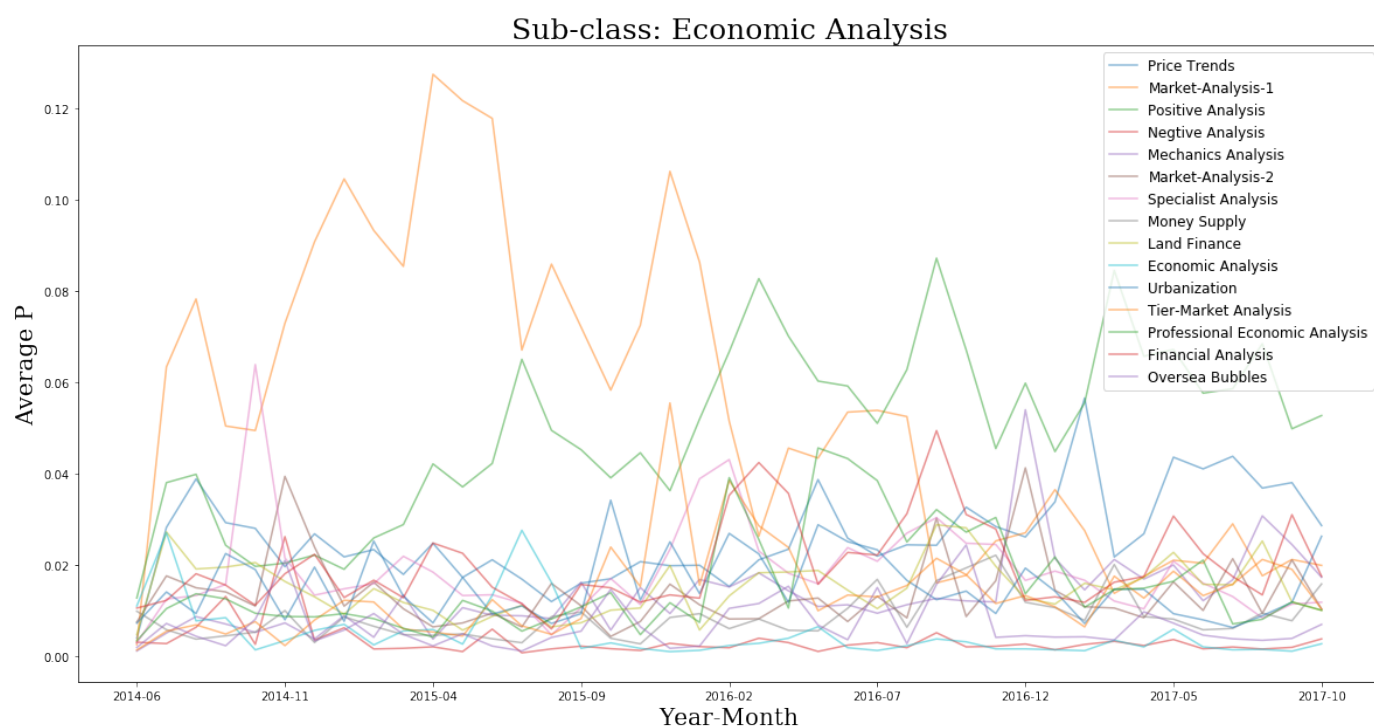
**Figure 5.2.2 Average Contribution Percentage of ‘Support’ Sub-Topics per Month**



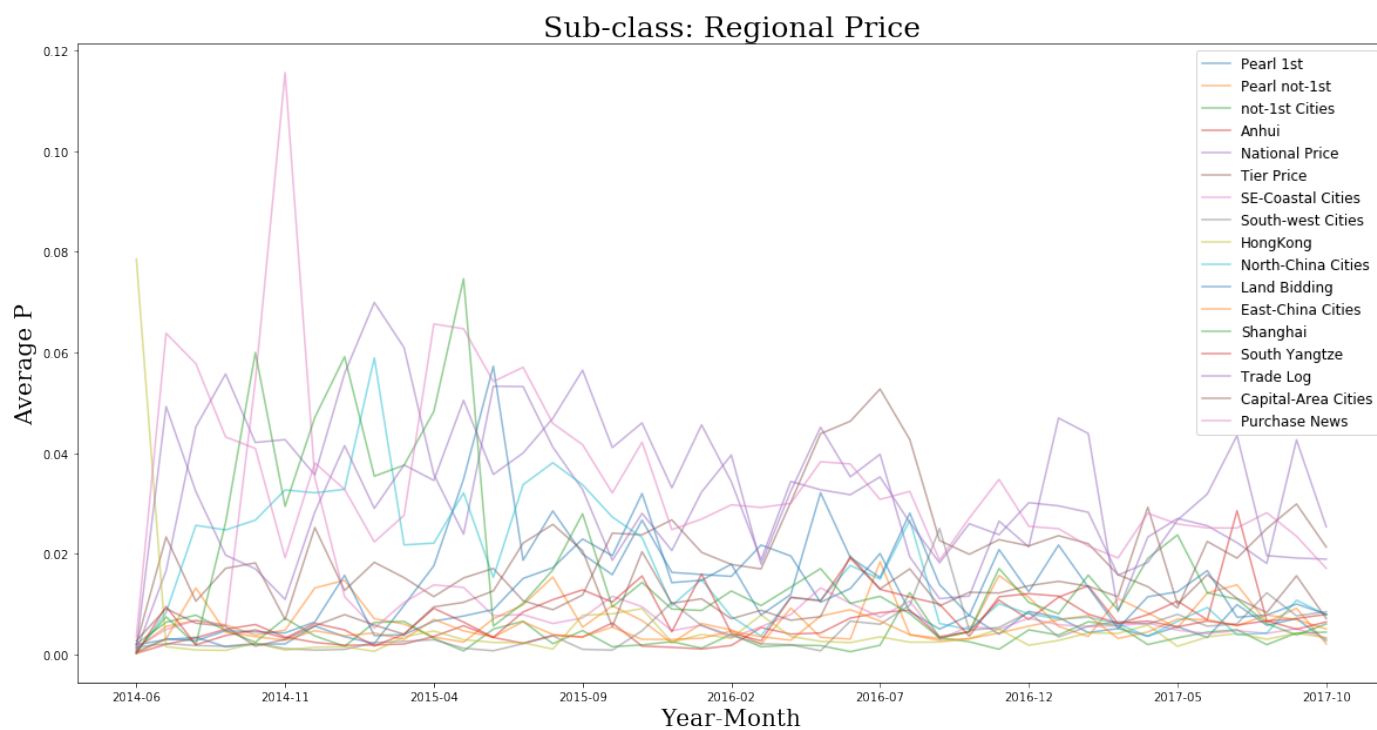
**Figure 5.2.3 Average Contribution Percentage of ‘Life Story’ Sub-Topics per Month**



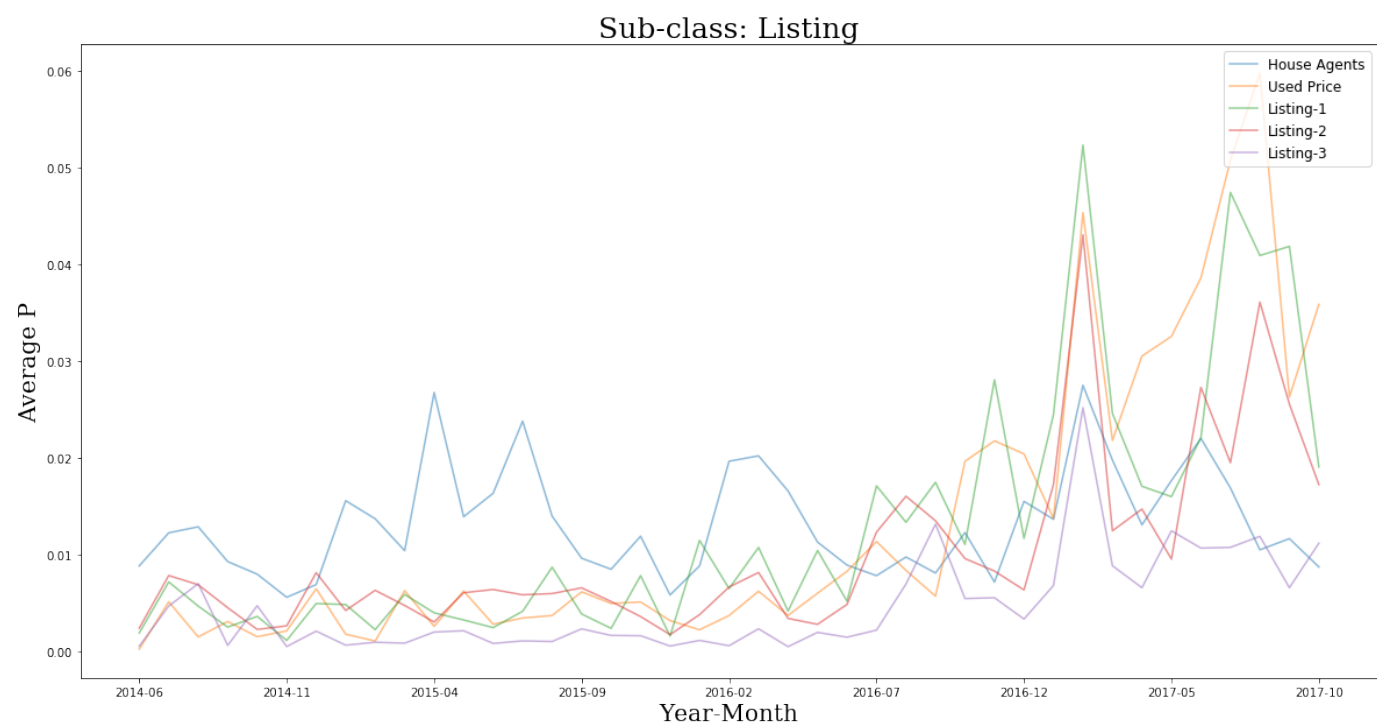
**Figure 5.2.4 Average Contribution Percentage of ‘Investment’ Sub-Topics per Month**



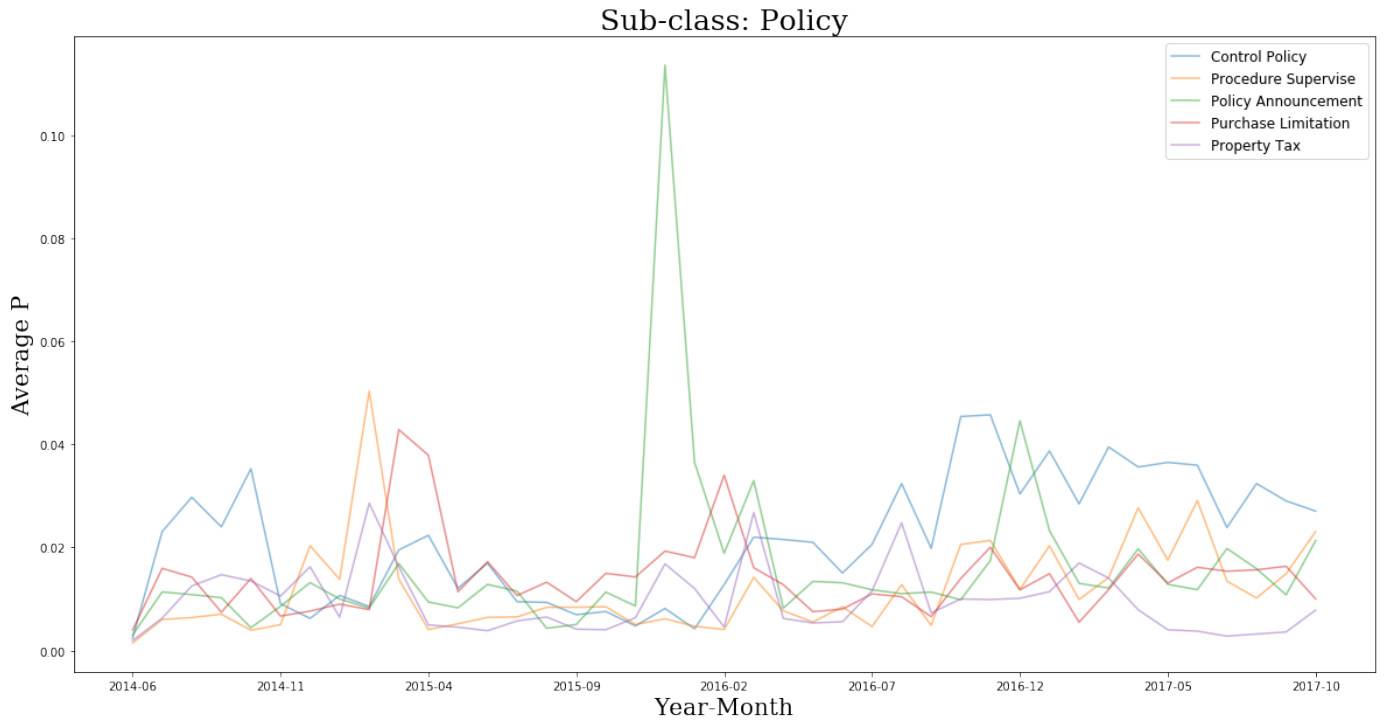
**Figure 5.2.5 Average Contribution Percentage of ‘Economic Analysis’ Sub-Topics per Month**



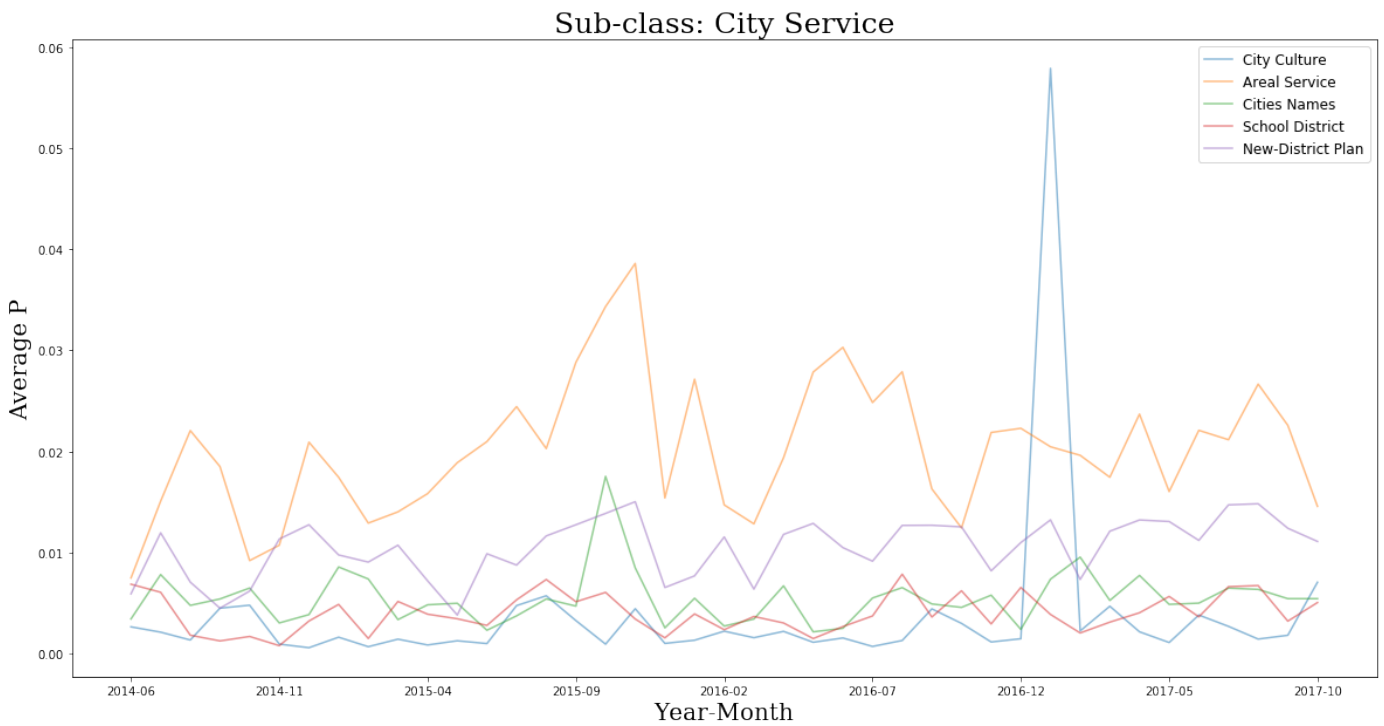
**Figure 5.2.6 Average Contribution Percentage of ‘Regional Price’ Sub-Topics per Month**



**Figure 5.2.7 Average Contribution Percentage of ‘Listing’ Sub-Topics per Month**



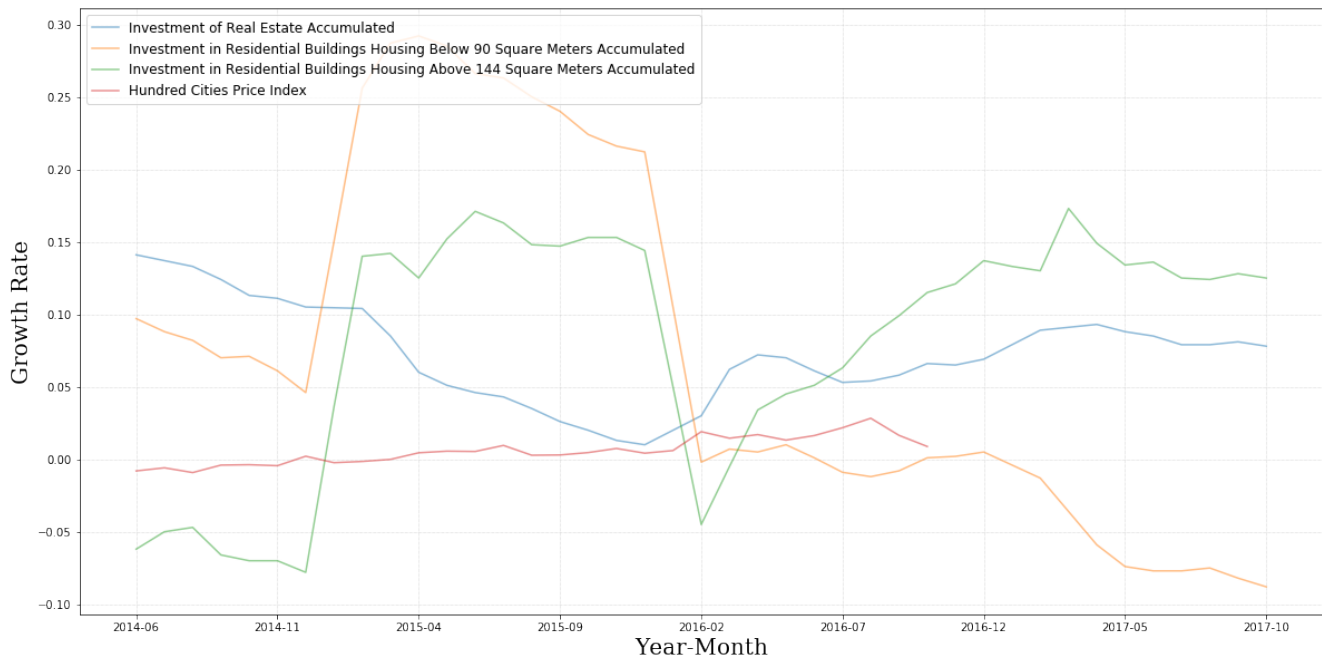
**Figure 5.2.8 Average Contribution Percentage of ‘Policy’ Sub-Topics per Month**



**Figure 5.2.9 Average Contribution Percentage of ‘City Service’ Sub-Topics per Month**

### **5.3 Real Estate Statistics**

In order to test out the association between main topics trends and real estate market situations, some statistics were collected that could reflect the trend of price and expectation of the market. To get an abstract insight of housing price, the Hundred Cities Price Index Growth Rate (from 2014-06 to 2016-10) from China Real Estate Index System of 'Fang.com' was used. The Hundred Cities Price Index is one of the important indicators of the real estate index and reflects the price trend of new homes sold at 100 significant cities at different times by sampling price in these cities. To measure the expectation of the market, the real estate investment statistics from National Bureau of Statistics of China was used because real estate developers are relatively sensitive to policy and market conditions. The exact statistics include total Investment in Residential Buildings in Real Estate Development Accumulated Growth Rate, Investment in Residential Buildings Housing Below 90 Square Meters Accumulated Growth Rate, and Investment in Residential Buildings Housing Above 144 Square Meters Accumulated Growth Rate.



**Figure 5.3.1 Real Estate Investment and Price Index per Month (2014-06~2017-10)**

All of the chosen statistics are about new homes. This is because new homes still compose significant amount of market sales nationwide in China since China is still undergoing a rapid urbanization process. Only very few first tier cities have considerable amount of second-hand house trades such as Beijing and Shanghai because they may have reached the optimum city size. Another thing that needs to be mentioned is that growth rates were used instead of original numbers for all statistics due to a preliminary assumption that market discussion is more sensitive to the change in growth than the change in original statistics. Line graphs of each growth rate is shown in the figure below. Additionally, in the table below, all the variables used and their explanations are listed in the following association analysis. Noted that the categorical variable ‘Event’ corresponds to different period of control policy (regulatory or stimulatory).

<b>Variable</b>	<b>Explanation</b>
<b>y</b>	Dependent variable for the regression model if it differs in each model in the table.
<b>C_Year</b>	Numeric variable indicating year but centered at 2014. Value includes 0, 1, 2, 3.
<b>Month</b>	Numeric variable indicating month. Value includes 1, 2, 3, ... , 10, 11, 12.
<b>P(Topic)</b>	Average contribution probability of a main topic per month. Value range: [0, 1].
<b>Event</b>	Categorical variable corresponds to major policy change. Values include 0, 1, 2. Value is 0 from 2014-06 to 2014-09, 1 from 2014-10 to 2016-11, 2 from 2016-12 to 2017-10. 0 and 2 represent regulatory policy period. 1 represents stimulus policy period.
<b>IREAG</b>	Investment of real estate accumulated growth rate by month (%).
<b>IRHB90AG</b>	Investment in residential buildings housing below 90 square meters accumulated growth rate by month (%).
<b>IRHA144AG</b>	Investment in residential buildings housing above 144 square meters accumulated growth rate by month (%).
<b>HCHPI</b>	Hundred cities price index growth rate by month (%).

**Table 5.3.1 Variable Explanations**

## **5.4 Association between Real Estate Statistics and Topics**

In order to identify topics that could contribute to policy decision in the future, it is necessary to reveal what topic is statistically associated to the market change. In this test section, different real estate statistics are used as the dependent variable and different “main topic contribution probability”, centered year, month, and the interaction term between contribution probability and centered year as the independent variables for the ordinary least square (OLS) regression model. The motivation behind these tests is to find significant associated independent variables, how they are associated, and how the slope changes with year. The null hypothesis is that no independent variable is significantly associated with the dependent variable.

A prediction model is not being built using the topic contribution as for the objective of this research, it is only necessary to see whether the statistical association between independent variable and dependent variables is significant or not. Although  $R^2$  is not important in this study, it is still included in the report. An interaction term between topic contribution probability and month was previously included in the model, however it turned out to be insignificant. This means the coefficient of topic contribution does not exhibit a significant monthly change so this interaction term was left out. In each of the following tables, one table represents one dependent variable and each row represent a different model that tested out the effect of the contribution probability of a certain main topic. No model involves more than one main topic.

In the most of following analysis of investment data, the centered year variable has a significant association, but month does not. This is because the investment data are very different from sales data which have a very obvious seasonal pattern. The difference between investment and sales is the inventory which does not have a seasonal pattern. In fact, Chinese government was focusing on reducing inventory in 2016, which affirms that inventory data is strongly variable.



**Investment of Real Estate Accumulated Growth rate (IREAG)**  
(2014-06 ~ 2017-10)

$$\text{IREAG} = \beta_0 + \beta_1 P(\text{Topic}) + \beta_2 C\_Year + \beta_3 \text{Month} + \beta_4 P(\text{Topic}) \times C\_Year$$

	Topic	(Intercept)	P(Topic)	C_Year	Month	P(Topic) × C_Year	R <sup>2</sup>	Adjusted R <sup>2</sup>
<b>Model_1</b>	<b>Complaint</b>	0.039573 (0.023793)	1.581071 (0.505408)**	0.019986 (0.010247)	-0.001395 (0.001509)	-0.928711 (0.338427)**	0.2896	0.2106
<b>Model_2</b>	<b>Support</b>	0.084813 (0.023951)**	0.498444 (0.675098)	-0.008175 (0.012690)	-0.001617 (0.001657)	0.024680 (0.518830)	0.1064	0.007106
<b>Model_3</b>	<b>Life Story</b>	0.092104 (0.018458)***	0.198953 (0.096071)*	-0.022431 (0.010523)*	-0.002299 (0.001507)	0.170334 (0.108838)	0.3047	0.2274
<b>Model_4</b>	<b>Investment</b>	0.045271 (0.040239)	0.531043 (0.249988)*	0.063812 (0.017364)***	-0.002932 (0.001355)*	-0.651104 (0.137288)***	0.4642	0.4047
<b>Model_5</b>	<b>Economic Analysis</b>	0.160675 (0.043636)***	-0.245859 (0.166997)	-0.007215 (0.029377)	-0.001372 (0.001588)	0.003832 (0.110453)	0.1928	0.1031
<b>Model_6</b>	<b>Regional Price</b>	0.156303 (0.039021)***	-0.165506 (0.150966)	0.003026 (0.021689)	-0.002560 (0.001554)	-0.086601 (0.099177)	0.27	0.1889
<b>Model_7</b>	<b>Listing</b>	0.123947 (0.024292)***	-0.253499 (0.804850)	-0.034294 (0.009041)***	-0.002402 (0.001465)	0.281459 (0.257350)	0.3256	0.2507
<b>Model_8</b>	<b>Policy</b>	0.146889 (0.026096)***	-0.703060 (0.334719)*	-0.055826 (0.017829)**	-0.001749 (0.001534)	0.608139 (0.223813)*	0.2364	0.1516
<b>Model_9</b>	<b>City Service</b>	0.1908611 (0.028864)***	-2.6565002 (0.719087)***	-0.0488887 (0.0136669)**	-0.0004769 (0.0014669)	1.0583446 (0.3150896)**	0.3322	0.258

Notes: Mean and standard errors shown in (parentheses). \*\*\* = p-value < 0.001, \*\* = p-value < 0.01, \* = p-value < 0.05.

**Table 5.4.1 Regression Analysis on IREA**

In the regression analysis on IREAG (Investment of real estate accumulated growth rate), an association was observed between IREAG and ‘Complaint’ topic contribution with a positive slope, and between IREAG and the ‘Complaint’ $\times$ Year interaction term with a negative slope. This means people complained about real estate market when there is growth in the market because it will lead to unequal distribution of wealth. Additionally, this effect is possibly reduced over the year because the government made some effort in alleviating this issue. Similarly, there is an association between IREAG and ‘Investment’ topic contribution with a positive slope, and between IREAG and the ‘Investment’ $\times$ Year interaction term with a negative slope. This is explainable since the dependent variable is also investment. The decrease in the interaction term, nonetheless, might indicate that investment return is becoming worse over the years.

There is association between IREAG and ‘Policy’ topic contribution with a negative slope, and between IREAG and the ‘Policy’ $\times$ Year interaction term with a positive slope. Since the slope of topic contribution is negative, this might reflect an interesting hypothesis that people discuss more during regulation policy period than stimulation policy period. The interaction being negative means there might be direction change in the real estate control policy. Similarly, it was noted that there is an association between IREAG and ‘City Service’ topic contribution with a negative slope, and between IREAG and the ‘City Service’ $\times$ Year interaction term with a positive slope. This might be because there is a lag between investment and urban infrastructure projects. This is also because in the real estate development, buildings with public service function (school, shopping mall) are usually built after the residential buildings.

**Hundred Cities House Price Index (HCHPI)**  
(2014-06 ~ 2016-10)

$$\text{HCHPI} = \beta_0 + \beta_1 P(\text{Topic}) + \beta_2 C\_Year + \beta_3 \text{Month} + \beta_4 P(\text{Topic}) \times C\_Year$$

	Topic	(Intercept)	P(Topic)	C_Year	Month	P(Topic) × C_Year	R <sup>2</sup>	Adjusted R <sup>2</sup>
<b>Model_1</b>	<b>Complaint</b>	-0.0131702 (0.0043275)**	0.0314169 (0.0841968)	0.0142788 (0.002667)***	0.0007411 (0.0002662)*	-0.0982440 (0.0786738)	0.8244	0.7951
<b>Model_2</b>	<b>Support</b>	-0.0148095 (0.003685)***	0.1074924 (0.0938137)	0.0136951 (0.002433)***	0.0007196 (0.0002693)*	-0.0758076 (0.0976332)	0.8195	0.7895
<b>Model_3</b>	<b>Life Story</b>	-0.0133303 (0.003516)***	0.0023007 (0.0147072)	0.0146170 (0.003077)***	0.0008137 (0.000291)**	-0.0425405 (0.0418566)	0.8178	0.7874
<b>Model_4</b>	<b>Investment</b>	-0.0063069 (0.0067371)	-0.0249058 (0.0403617)	0.0172982 (0.004334)***	0.0004885 (0.0002492)	-0.0445704 (0.0297046)	0.8609	0.8377
<b>Model_5</b>	<b>Economic Analysis</b>	-0.0112030 (0.0062244)	-0.0012403 (0.0238316)	0.0087127 (0.0065084)	0.0006921 (0.0002846)	0.0093126 (0.0205938)	0.8117	0.7803
<b>Model_6</b>	<b>Regional Price</b>	-0.0133036 (0.0057969)*	0.0002703 (0.0226186)	0.0076288 (0.0042791)	0.0007963 (0.000269)**	0.0191197 (0.0181982)	0.8318	0.8038
<b>Model_7</b>	<b>Listing</b>	-0.0127135 (0.0054859)*	0.0749658 (0.1631217)	0.0101352 (0.0035304)**	0.0006269 (0.0002930)*	0.0184428 (0.1005908)	0.8203	0.7903
<b>Model_8</b>	<b>Policy</b>	-0.0108147 (0.0043533)*	-0.0207463 (0.0597823)	0.0120731 (0.0033789)**	0.0007180 (0.0002735)*	-0.0011459 (0.0498888)	0.8144	0.7835
<b>Model_9</b>	<b>City Service</b>	-0.0074694 (0.0064843)	-0.1086386 (0.1612797)	0.0067508 (0.0057557)	0.0006440 (0.0002969)*	0.1268266 (0.1435199)	0.816	0.7853

Notes: Mean and standard errors shown in (parentheses). \*\*\* = p-value < 0.001, \*\* = p-value < 0.01, \* = p-value < 0.05.

**Table 5.4.2 Regression Analysis on HCHPI**

In the regression analysis on HCHPI (hundred cities price index growth rate by month), no topic contribution variable is significantly associated. Since HCHPI is an averaging index sampling from 100 cities and each city has a different growth rate from others, HCHPI is more like a general indication of a long term price trends. Considering this fact, it is possible to conclude that online discussion of real estate focuses more on things that happen currently or in the near future. Topic contribution is more helpful for short-term analysis than for long-term analysis.

**Investment in Residential buildings Housing Below 90 square meters Accumulated Growth rate (IRHB90AG)**  
(2014-06 ~ 2017-10)

$$IRHB90AG = \beta_0 + \beta_1 P(Topic) + \beta_2 C\_Year + \beta_3 Month + \beta_4 P(Topic) \times C\_Year$$

	Topic	(Intercept)	P(Topic)	C_Year	Month	P(Topic) × C_Year	R <sup>2</sup>	Adjusted R <sup>2</sup>
<b>Model_1</b>	<b>Complaint</b>	0.404901 (0.067728)***	-3.486062 (1.438680)*	-0.160672 (0.029168)***	-0.010577 (0.004296)*	2.624826 (0.963356)**	0.5797	0.533
<b>Model_2</b>	<b>Support</b>	0.318019 (0.066530)***	-1.778067 (1.875282)	-0.112794 (0.035251)**	-0.009143 (0.004602)	1.104942 (1.441203)	0.4966	0.4407
<b>Model_3</b>	<b>Life Story</b>	0.310941 (0.054814)***	-0.559410 (0.285300)	-0.080427 (0.031251)*	-0.008823 (0.004475)	-0.101743 (0.323216)	0.5523	0.5026
<b>Model_4</b>	<b>Investment</b>	0.427629 (0.139890)**	-1.185204 (0.869090)	-0.192062 (0.060367)**	-0.008755 (0.004710)	0.867748 (0.477284)	0.5272	0.4747
<b>Model_5</b>	<b>Economic Analysis</b>	0.12996 (0.12393)	0.60968 (0.47428)	0.05823 (0.08343)	-0.01009 (0.00451)*	-0.55122 (0.31369)	0.5246	0.4718
<b>Model_6</b>	<b>Regional Price</b>	0.040760 (0.108112)	0.804791 (0.418266)	-0.063303 (0.060093)	-0.007760 (0.004305)	-0.021820 (0.274780)	0.5909	0.5454
<b>Model_7</b>	<b>Listing</b>	0.220042 (0.076630)**	1.347410 (2.538954)	-0.052706 (0.028519)	-0.008674 (0.004621)	-0.646985 (0.811828)	0.5101	0.4556
<b>Model_8</b>	<b>Policy</b>	0.166738 (0.076055)*	1.624040 (0.975514)	-0.000589 (0.051961)	-0.009313 (0.004472)*	-1.156300 (0.652286)	0.5265	0.4739
<b>Model_9</b>	<b>City Service</b>	0.072316 (0.085573)	5.786783 (2.131914)*	-0.005259 (0.040519)	-0.011809 (0.004349)*	-2.189561 (0.934162)*	0.5715	0.5238

Notes: Mean and standard errors shown in (parentheses). \*\*\* = p-value < 0.001, \*\* = p-value < 0.01, \* = p-value < 0.05.

**Table 5.4.3 Regression Analysis on IRHB90A**

In the regression analysis on IRHB90AG (Investment in residential buildings housing below 90 square meters accumulated growth rate), an association was observed between IRHB90AG and ‘Complaint’ topic contribution with a negative slope, and between IRHB90AG and ‘Complaint’ $\times$ Year interaction term with a positive slope. Since a residential unit below 90 square meters is usually from an affordable housing project, the slope of ‘Complaint’ $\times$ Year interaction term grows over the years because the Chinese government has recently started to dedicate itself to balancing wealth distribution and affordable housing. The reason for the ‘Complaint’ slope being initially negative might be that people ‘Complaint’ less under the stimulus policy time. There is an association between IRHB90AG and ‘City Service’ topic contribution with a positive slope, and between IRHB90AG and ‘City Service’ $\times$ Year interaction term with a negative slope. This might be because the poor city service ability is one of the reasons for people’s complaints, and with the government’s effort, this issue is alleviated over time.

The regression analysis table was excluded for IRHA144AG (Investment in residential buildings housing above 144 square meters accumulated growth rate) because no independent variable including year and month is significant in any model. Residential units over 144 square meters are either high-end properties or are designed for big families. A possible reason for insignificant association might be due to WeChat articles failing to capture these groups of readers.

## 5.5 Association between Control Policy and Topics

After testing the association between topics and market, it is possible to see whether the topic contributions are associated with a specific policy event. As the purpose for this study is to find whether it is possible for governments to make practical policy changes to solve market problems using online discussion as a tool. If the topic contributions are sensitive to different policy conditions, online discussion may be used as an intermediary to effectively connect market change with policy change.

The tool used for testing topic-policy sensitivity is the R Emmeans package which estimates marginal means (least-squares means). Emmeans computes estimated marginal means for a specified factor variable in a linear model and gives a comparison among them. The independent variable included are “Centered Year” and “Event”. Month was excluded since it is already known that topic contribution does not have a seasonal trend. The dependent variable is different in each row. The first row is IREAG (Investment of real estate accumulated growth rate) and the rest of nine rows are the nine main topics. Then using this linear model as the input, the estimated margin means analysis is performed to see if the dependent variable is significantly different in slope of “Event” in a pair-wise period comparison. In this test, the “Event” factor is a dummy variable. Its value is 0 from 2014-06 to 2014-09, representing a regulatory policy period. Its value is 1 from 2014-10 to 2016-11, representing a stimulatory policy period. Its value is 2 from 2016-12 to 2017-10, representing a regulatory policy period.

**OLS Regression Analysis and Estimated Margin Means (over Event)**  
(2014-06 ~ 2017-10)

$$y = \beta_0 + \beta_1 C\_Year + \beta_2 Event \text{ (Event is a factor variable [0 / 1 / 2])}$$

	OLS Regression					Estimated Margin Means		
	<i>y</i>	(Intercept)	C_Year	Event-1	Event-2	0-1	0-2	1-2
<b>Model_0</b>	<b>IREAG</b>	0.133750 (0.012094)***	-0.014240 (0.006856) *	-0.055993 (0.015785)**	-0.009505 (0.024439)	0.055993258 (0.01578454)*	0.009505056 (0.02443851)	-0.0464882 (0.0140082)*
<b>Model_1</b>	<b>Complaint</b>	0.049569 (0.006026)***	-0.006520 (0.003416) +	-0.014125 (0.007865) +	-0.012653 (0.012177)	0.014124812 (0.007865043)	0.012652870 (0.012177103)	-0.001471943 (0.006979972) )
<b>Model_2</b>	<b>Support</b>	0.0269211 (0.0056417)***	-0.0038741 (0.0031982)	0.0004216 (0.0073632)	0.0048941 (0.0114001)	-0.0004216472 (0.007363183)	-0.004894134 (0.011400094)	-0.0044724873 (0.006534587)
<b>Model_3</b>	<b>Life Story</b>	0.149911 (0.023669)***	0.003029 (0.013417)	-0.100660 (0.030891)**	-0.079060 (0.047827)	0.10066013 (0.03089105)*	0.07906043 (0.04782726)	-0.02159970 (0.02741481)
<b>Model_4</b>	<b>Investment</b>	0.171197 (0.011296)***	0.002403 (0.006403)	-0.040058 (0.014743)**	-0.077713 (0.022826)**	0.04005811 (0.01474290) +	0.07771331 (0.02282578)*	0.03765521 (0.01308385) +
<b>Model_5</b>	<b>Economic Analysis</b>	0.22273 (0.02076)***	0.02846 (0.01177)*	0.03433 (0.02710)	-0.03772 (0.04195)	-0.03433149 (0.02709568)	0.03772349 (0.04195106)	0.07205498 (0.02404654) +
<b>Model_6</b>	<b>Regional Price</b>	0.19484 (0.02193)***	-0.03769 (0.01243)**	0.11939 (0.02862)***	0.11402 (0.04431)*	-0.119387838 (0.02862166)***	-0.114024091 (0.04431367) +	0.005363747 (0.02540080)
<b>Model_7</b>	<b>Listing</b>	0.026245 (0.011224)*	0.016840 (0.006363)*	-0.013447 (0.014649)	0.035283 (0.022681)	0.01344655 (0.01464917)	-0.03528301 (0.02268067)	-0.04872956 (0.01300066)
<b>Model_8</b>	<b>Policy</b>	0.053216 (0.013876)***	0.006654 (0.007866)	0.008167 (0.018110)	0.017492 (0.028039)	-0.008167428 (0.01811011)	-0.017491856 (0.02803909)	-0.009324427 (0.01607214)
<b>Model_9</b>	<b>City Service</b>	0.035139 (0.006742)***	0.003112 (0.003822)	0.002805 (0.008799)	0.007025 (0.013623)	-0.002804661 (0.008798820)	-0.007025013 (0.013622828)	-0.004220351 (0.007808668)

Notes: Mean and standard errors shown in (parentheses). \*\*\* = p-value < 0.001, \*\* = p-value < 0.01, \* = p-value < 0.05, + = p-value < 0.1.

Notes: IREAG = Investment of Real Estate Accumulated Growth rate

**Table 5.5.1 Regression Analysis and Emmeans Analysis**



The reason for including the IREAG test is to assure the market investment is actually statistically different during transition periods (0 to 1, and 1 to 2), which turn out to be true. The ‘Life Story’ topic is significantly different from period 0 to 1 with a decreasing slope. The ‘Investment’ topic is weakly significantly different from period 0 to 1 and from period 1 to 2 with a decreasing slope. The ‘Economic Analysis’ topic is weakly significantly different from period 1 to 2. The ‘Regional Price’ topic is strongly significantly different from period 0 to 1 with an increasing slope. The reason for the greater topic difference between period 0 and period 1 than period 1 and period 2 might be due to the online discussion being more sensitive to the stimulatory policy change than the regulatory policy change. This might need some time for the market to cool down over discussion of market expansion.

## **6 Time Series Analysis - Granger Causality Test**

From the regression analysis above, it can be concluded that there is some significant relationship or topic trend influencing market conditions. The correlation between topics and market is an important finding that might help government in making guiding policy to steer the direction of market trends. In the previous section, regression was used as a static model representing the relationship between real estate market and online discussion at given points in time. Dynamic models can also be used to analyze time-dependent changes of this relationship. From macroeconomics studies in targeting inflation rate, it is beneficial for governments to control the inflation target to improve economic stability. (Friedman, 2008) A similar effect can be assumed on the price estimation. If one can test the causal relationship between topics and market

trend or vice versa, it will be very helpful to adjust controlling policy with respect to this causal relationship. Further inference can be made about the causal relationship using time series analysis.

The Granger causality test is a statistical test on the concept of causality based on the effectiveness in prediction. While ordinary regression tests reflect merely the correlation of two variables in separate time, the Granger causality test can tell if the value of one time-series value is significantly helpful in predicting another. Since these data are temporal records, one can determine how to effectively use historical records to predict topic trends or market trends in the future.

Granger defines Granger causality as: ‘A time series variable X Granger causes Y, if the probability of Y conditional on its own past history and the past history of X does not equal the probability of Y conditional on its own past history alone.’ Since the question of real causality is deeply philosophical, the Granger causality test is considered to give only predictive causality. From an empirical view of controlling the market, however, predictive causality is a good indicator for policy-makers to figure out what factors to focus on in the complicating markets. For now, the direction of causality is unknown, so there may be different possible hypotheses.

## **6.1 Test Hypothesis**

In previous economic research, a hypothesis can also be made to test a null hypothesis in a Granger causing relationship. (Lee, Lin, Chuang & Lee, 2011) The first hypothesis is that the online topics trends are useful in forecasting market trends, but not vice versa. This hypothesis assumes a situation in which online discussion has a direct impact on the market condition. If this is so, online discussion is possibly reflecting the purchasing will and the purchasing power in the real estate buyer, and such factors will affect the real estate market in the next time period. It is

also possible that government has been taking advice from the online discussion and been making corresponding policy changes to alleviate the problems which are discussed in the previous time period. If the online topic trends are not useful in forecasting market trends, it is possible that online discussion does not reflect people's decision in purchasing houses.

The second hypothesis is that the market trends are useful in forecasting the online topics trends, but not vice versa. This hypothesis imagines a situation where the market trends affect the online real estate topics discussed in the following period. If the market trends are not useful in forecasting the topic trends, the online discussions are probably not sensitive to the market change.

The third hypothesis is that the previous hypotheses are both valid. Under this hypothesis, the topics trends are useful in predicting the market trend and the market trend is useful in predicting the topics trends. This hypothesis is more likely to be true because there is usually no absolute, single-direction causal relationship. If this hypothesis is valid, it could mean that market condition has influence on the topics to be discussed in the future while the current discussion could also lead to different market trends in the future. If neither of the hypotheses are true, one might need to reconsider whether there is a significant correlation between online topics and the real estate market.

## **6.2 Data Preprocessing**

To prepare the current data for the Granger causality test, the topics percentage data was differenced by subtracting the previous percentage and getting the marginal growth of the topics percentage. Since the market statistics are already in growth rate form, it is not necessary to further transform them. The reason for differencing the data is because there is a prior assumption that the online discussion is mostly sensitive to the change in the market and vice versa, since in the entire

pool of market discussion there are always different kinds of opinions on different topics. When there are changes in the proportion of a topic, it could mean this topic became more or less popular and could cause a change in the market.

Another important reason for differencing the time series data is to avoid non-stationarity. (Chiou-Wei, Chen & Zhu, 2008) Granger and Newbold discovered that it is possible to get spurious regression when analyzing non-stationary time series data, which could mistakenly make two unrelated variables causally related. It is therefore necessary to make the topics and market trends data stationary. To make sure data is stationary, the Dickey–Fuller test was used on each of the time series variables after differencing from first order to a higher level difference order until an order could make this time series variable pass the Dickey–Fuller test.

### **6.3 Result Analysis**

In the Granger causality test of this study, only the Granger causing relationship was tested between topics and Investment of Real Estate Accumulated Growth rate (IREAG). This is because only this statistic contains information of enough complexity that can generally reflect an overall trend of the real estate market. Other statistics are either too specific or cover too short of a time span. The lag parameters of VAR, the vector auto-regression model that is used for the causality test, is selected based on choosing the model having the lowest AIC so that the model can contain an appropriate time lag. The null hypothesis column represents the null hypothesis to be tested. For example, the row with ‘Complaint  $\nRightarrow$  IREAG’ represents the null hypothesis: the variable topic “Complaint” does not Granger cause variable IREAG.

**Granger Causality Results (2014-06 ~ 2017-10)**  
Investment of Real Estate Accumulated Growth rate (IREAG)

	Topics	IREAG	Lags	Null Hypothesis	F-Statistics	P-Value
Complaint	I(1)	I(1)	9	Complaint $\neq$ IREAG	2.8814	<b>0.0508 †</b>
				IREAG $\neq$ Complaint	0.3797	0.9216
Support	I(1)	I(1)	8	Support $\neq$ IREAG	2.5998	<b>0.0565 †</b>
				IREAG $\neq$ Support	1.9745	0.1269
Life Story	I(2)	I(1)	9	Life Story $\neq$ IREAG	0.8932	0.5626
				IREAG $\neq$ Life Story	1.1100	0.4333
Investment	I(1)	I(1)	9	Investment $\neq$ IREAG	2.4491	<b>0.0817 †</b>
				IREAG $\neq$ Investment	1.8033	0.1766
Economic Analysis	I(1)	I(1)	2	Economic Analysis $\neq$ IREAG	6.1580	<b>0.0055 **</b>
				IREAG $\neq$ Economic Analysis	1.3839	0.2652
Regional Price	I(1)	I(1)	8	Regional Price $\neq$ IREAG	3.0222	<b>0.0339 *</b>
				IREAG $\neq$ Regional Price	1.1731	0.3790
Listing	I(8)	I(1)	1	Listing $\neq$ IREAG	0.0004	0.9845
				IREAG $\neq$ Listing	0.0003	0.9859
Policy	I(1)	I(1)	1	Policy $\neq$ IREAG	0.0206	0.8867
				IREAG $\neq$ Policy	0.0580	0.8111
City Service	I(1)	I(1)	1	City Service $\neq$ IREAG	0.1530	0.6980
				IREAG $\neq$ City Service	0.0231	0.8800

Notes:

Mean and standard errors shown in (parentheses). \*\*\* = p-value < 0.001, \*\* = p-value < 0.01, \* = p-value < 0.05, † = p-value < 0.1.

**Table 6.3.1 Granger Causality Test Result**

The topics of ‘Complaint’, ‘Support’, ‘Investment’, ‘Economics Analysis’, and ‘Regional Price’ have a relatively high probability to associate with Real Estate Investment in the Granger causing relationship. While other topics like ‘Life Story’, ‘Listing’, ‘Policy’, ‘City Service’ do not

have such a high probability to deny the null hypothesis. This result makes sense because the Granger causing topics are directly related to different expectation in investment. One can imagine a loosening policy stage when real estate price is soaring versus a tightening policy stage when price growth slows down; online discussion here should have different proportion in these Granger causing topics. Other topics tend to be more stable despite the market condition.

Another important observation is that, in the test result, only the topics variables Granger cause market investment growth rate, but not vice versa. This means the first hypothesis, that online topic trends are useful in forecasting market trends, is more likely to be correct. However, the market investment growth does not Granger cause change in online discussion. It can be inferred that the online discussion of real estate market is more about forecasting the future, it has the power of influencing buyer's and developer's decisions. Since investment has a relatively long-term return, it probably cannot have short term effect on topics. Meanwhile, topics probably can instantly influence people's expectation of market trends. This is very significant information for policy makers because it serves as an indicator for future market trends. Policy makers can either utilize such information to adjust policy accordingly or make announcements to shift the online discussion which might cause change in market expectation.

## **7 Conclusion**

In this study, a complete research pipeline was gone over including collecting data, processing data, re-structuring data, and analyzing data. The approach of text mining unearths some degree of inference ability for online discussion of the real estate market. This study can be seen as foundational for research in the social sciences for further exploration. Through the result of data analysis, one can observe what different topic discussions people will have under different

market and policy situations. By time series analysis and the Granger causality test, one is more clear about how online discussion is related to the market. Discussion topics have a lot of potential in revealing buyer's expectation in the future. The policy-maker may use this approach to achieve the social effect they desire to realize a better market mechanism design. Although being a relatively large-scale data analysis study, it is possible that the relatively small amount of data could have restricted the inference ability of the text data. It is still necessary to be very careful when using text mining approaches because there still might be some problems with biased data and biased user groups. Additional qualitative research on online users and the market need to be done to assure that the current inferences are reasonable.

Despite the limitations of this study, it has implications for the market situation, topics discussion and controlling policy. It is an exploratory methodology study that exploits the data that is not available but will be trending in the future. Especially in the digital age, online discussion will be more and more related to the real world. With the online information distributing effect like echo chamber, it is more than necessary to take online information into modeling consideration. Meanwhile, the text mining tool used in this study turns out to be a powerful tool to aid with social science studies. It provides an efficient access to public opinions that traditionally requires a lot of monetary and human resources. We can see the potential in data to support behavioral studies in real estate policy such as behavioral finance or behavioral economics. The amount of online text data will keep growing and penetrate more and more user groups which means the inferences drawn from these data will very likely become more and more reliable.

Finally, several areas in this study could be expanded for further research. First, each WeChat article has some affiliate data that have not been used in this study, for example, the information of the publishing account, number of likes and views, and the comments from the

readers. These data can be used to strengthen the measurement of topic contributions and their influence over readers. Second, more sophisticated time series analyses and causality tests could be used to help policy-makers in making successful decisions. Usually, time series analyses need more data. We can more confidently test time variant models in the future. Third, this research approach could be combined with a behavioral study or experiment to test out the psychological effects of the real estate market on people. Fourth, China is developing a policy strategy where each city should follow a separated control guideline which means the market will be largely segmented. Following the study, more analysis can be done in sub markets or sub geographical regions.



## BIBLIOGRAPHY

Du, Z., & Zhang, L. (2015). Home-purchase restriction, property tax and housing price in China: A counterfactual analysis. *Journal of Econometrics*, 188(2), 558-568. doi:10.1016/j.jeconom.2015.03.018

Wang, W., & Ye, F. (2016). The Political Economy of Land Finance in China. *Public Budgeting & Finance*, 36(2), 91-110. doi:10.1111/pbaf.12086

Huang, G., & Cai, J. (2013). The Root of "Land Finance": Tax System and Countermeasures. *Macro Economic Research*.

Tian, Q., & Zhu, H. (2016). Discussion on China's housing prices. *Southern Finance*.

Qun, W., Yongle, L., & Siqu, Y. (2015). The incentives of Chinas urban land finance. *Land Use Policy*, 42, 432-442. doi:10.1016/j.landusepol.2014.08.015

Cao, G., Feng, C., & Tao, R. (2008). Local "Land Finance" in Chinas Urban Expansion: Challenges and Solutions. *China & World Economy*, 16(2), 19-30. doi:10.1111/j.1749-124x.2008.00104.x

Zhao, Y., & Webster, C. (2011). Land Dispossession and Enrichment in China's Suburban Villages. *Urban Studies*, 48(3), 529-551. doi:10.1177/0042098010390238

Yang, Z., & Chen, J. (2014). Housing Reform and the Housing Market in Urban China. *SpringerBriefs in Economics Housing Affordability and Housing Policy in Urban China*, 15-43. doi:10.1007/978-3-642-54044-8\_2

Zhao, Y. (2016). National Credit and Land Finance --- China's urbanization facing transformation. *Urban Economics*, 1006(3862).

Wang, R., Hou, J., & He, X. (2017). Real estate price and heterogeneous investment behavior in China. *Economic Modelling*, 60, 271-280. doi:10.1016/j.econmod.2016.09.020

Xiang, F. (2017). Government Credit Construction in the Network Age. *Policy and Commercial Law Research*.

Zhao, C., & Shen, G. (Eds.). (2018, July 23). 2018 Statistical Report on the Development of China's Internet Network. Retrieved from <http://tc.people.com.cn/n1/2018/0723/c183008-30164524.html>

- Goerge, R., Ozik, J., & Collier, N. (2015). Bringing big data into public policy research: Text mining to acquire richer data on program participants, their behavior and services. *Association for Public Policy Analysis and Management, Big Data and Public Policy Workshop*.
- Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet allocation. *Machine Learning Research*.
- Ko, N., Jeong, B., Choi, S., & Yoon, J. (2018). Identifying Product Opportunities Using Social Media Mining: Application of Topic Modeling and Chance Discovery Theory. *IEEE Access*, 6, 1680-1693. doi:10.1109/access.2017.2780046
- Ohsawa, Y. (2003). KeyGraph: Visualized Structure Among Event Clusters. *Chance Discovery*, 262-275. doi:10.1007/978-3-662-06230-2\_18
- Han, H. S., Mankad, S., Gavirneni, N., & Verma, R. (2016). What Guests Really Think of Your Hotel: Text Analytics of Online Customer Reviews. *Cornell University Hospitality Administration and Management Commons*.
- Lee, L., Lin, P., Chuang, Y., & Lee, Y. (2011). Research output and economic productivity: A Granger causality test. *Scientometrics*, 89(2), 465-478. doi:10.1007/s11192-011-0476-9
- Chiou-Wei, S. Z., Chen, C., & Zhu, Z. (2008). Economic growth and energy consumption revisited — Evidence from linear and nonlinear Granger causality. *Energy Economics*, 30(6), 3063-3076. doi:10.1016/j.eneco.2008.02.002
- Friedman, B. M. (2008). Monetary policy for emerging market economies: Beyond inflation targeting. *Macroeconomics and Finance in Emerging Market Economies*, 1(1), 1-12. doi:10.1080/17520840801903083

# APPENDIX

Topic Name	ID	Rank_0	Rank_1	Rank_2	Rank_3	Rank_4	Rank_5	Rank_6	Rank_7	Rank_8	Rank_9
Company Issues	0	Company	Yuan	Wenzhou	Enterprise	Corporate	Developer	Money	Funding	Contract	Home
Construction	1	area	building	square meters	floor	plan	engineering	flat meter	residential	total	high
China Development	2	China	Economy	Development	Social	Future	Becoming	Corporate	Problem	Country	Wealth
Pearl 1st	3	Shenzhen	Housing Price	Wan	Dongguan	Shenzhen	Longhua	Huizhou	Yuan	District	Deep
Pearl not-1st	4	yuan	average price	room	address	zhuhai	boulevard	discount	sale	property	interchange
Price Rank	5	house price	ranking	cities	rank	city	nation	No.	average price	China	county
House Agents	6	intermediary	second-hand housing	price	million	customer	owner	house	residential	sale	room
Company News	7	real estate	real estate	company	corporation	enterprise	project	industry	vanke	group	yuan
Price Tease	8	House	Buy	Million	Money	Sell	Dollar	Set	RMB	Price	Approx
Price Trends	9	House Prices	Down falling	Falling	Property Market	Uprising	Market	Price Drop	Rising	Appearing	Already
Travel Island	10	Hotel	Island	Eat	Yuan	Taiwan	Lane	Recommended	Travel	Taipei	Beach
Price Complaint	11	media	buy	markdown	woman	set	into	no	this	government	people
Market-Analysis-1	12	Market	City	Property Market	Policy	Real Estate	House Prices	Inventory	Rising	Demand	First Line
Mortgage	13	Bank	Interest Rate	Loan	Mortgage	Dollar	Interest	Benchmark	Repayment	First	Ten thousand
Used Price	14	yuan	average price	second-hand housing	housing prices	rose	last month	ring	real estate	new home	residential
Listing-1	15	Garden	Community	International	Plaza	New	Residence	City	Homes	Apartment	New Village
not-1st Cities	16	Zhengzhou	Brand	Henan	Zhengdong	New District	Zhengzhou	square	Company	Consumer	Market
Positive Analysis	17	buy	house	house price	buy	room	set	now	up	no	thinking
Anhui	18	Hefei	housing prices	yuan	Binhu	Anhui	Wan	Weifang	blocks	units	price
Increasing	19	House prices	Real estate	Rising	City	Room	China	Buy house	Development	Regulation	House
Service Industry	20	Hotel	Guest	Service	Customer	Expense	Management	Product	Business	Use	Date
Negative Analysis	21	House	Rose	China	Drop	Real Estate	Crash	Country	Buy	Hong Kong	Now
National Price	22	City	Price	Dwelling	Rising	Rising	Rate	Mom	Drop	Year	New
Mechanics Analysis	23	house prices	urban	land	income	shanghai	rise	rise	supply	beijing	house
Control Policy	24	Policy	Regulation	Property Market	Restriction Purchase	Promulgation	Market	Real Estate	Set	First	Limiting Loan
Life Plan	25	Buy	Money	Work	Kids	Life	No	Two	Parents	Living	University
Anecdote	26	The Last	Dial	Pan	The Age	Ishigaki	No	Economics	Inside	Into	Batch
Procedure Supervise	27	Registration	Real estate	Real estate	Sector	Sales	Marketing	Pre-sale	Commercial house	Information	Price

Headline Trash-News	28	China	Became	Beijing	First	World	Network	No	Times	Wukong	Eat
Market-Analysis-2	29	Now	China	No	Market	Money	Growth	Possible	Problems	House	Economy
Contract Dispute	30	Contract	Houses	Breaching	Mr.	Two sides	Meta	Sale	Signing	Request	Seller
City Culture	31	Beijing	Many	China	Culture	Shao	National	Shanghai	Men	No	Out
Other Price	32	Jade	Kind	Yuan	Value	Million	Glass	Time	Law	Garlic	University
Tier Price	33	Cities	Housing Prices	Shanghai	Shenzhen	Nanjing	Rising	Beijing	Guangzhou	Hangzhou	Xiamen
SE-Coastal Cities	34	Xiamen	housing prices	Tencent	real estate	housing	Fuzhou	yuan	United Network	news	buy a house
South-west Cities	35	Chengdu	House Prices	Million	Shanghai	Tianfu	Guiyang	Circle	Kilometer	Strip	Plate
Vulgar Story-1	36	A	teacher	Zhang Ergou	junior high school	dog	education	provincial capital	study	no	inside
Salary Expense	37	Average	Salary	Chongqing	House Prices	China	January	Eat	Drink	Beijing	Rent
Hong Kong	38	Hong Kong	Balance	Mainland	Think	China	Times	Block	House Prices	Hong Kong Dollar	United States
Travel Domestic	39	Features	Yuan	Sanya	Dali	Travel	Hainan	Lijiang	West Lake	Accommodation	Events
Specialist Analysis	40	house price	real estate	china	bubble	market	economy	rising	cycle	policy	rising
Listing-2	41	yuan	square meters	residential	garden yuan	garden	city yuan	average price	international	house prices	homes
Policy Announcement	42	Development	Real estate	Economy	Market	Reform	Policy	Housing	Work	Meeting	Central
Money Supply	43	RMB	China	Dollar	Economy	Exchange Rate	Devaluation	United States	Global	Assets	Interest Rate Increase
North-China Cities	44	residential	ring	drop	price	rise	square meter	city	index	house price	average price
Oversea Market	45	House Prices	Australia	Sydney	Canada	Housing	Real Estate	Vancouver	Wanwan	District	Overseas
Poor-Story-1	46	elderly	China	aged	selling	bit	money	buy	out	kids	no
Land Finance	47	Land	Government	Real Estate	House Prices	China	Local	Problem	Urban	Financial	Revenue
Economic Analysis	48	China	money	house price	price	economy	developer	government	no	island	exchange
Land Bidding	49	Suzhou	Land	Land	Yuan	Land price	Diwang	Wan	Song	Concession	Landscape
East-China Cities	50	Hangzhou	Jiaxing	Plate	Zhejiang	City	Xiaoshan	Binjiang	Qianjiang	Property	Apartment
Urbanization	51	Population	City	Million	Growth	Region	Inflow	Housing price	Industry	Occupation	Development
Poor-Story-2	52	Yang Pei	mother	child	no	cross stitch	embroidery	works	like	Shanghai	friends
Areal Service	53	area	matching	subway	line	center	traffic	project	commercial	planning	life
Listing-3	54	Yuan	community	home	garden	homeland	Tianjin	new	apartment	real estate	room price
Shanghai	55	Shanghai	Yuan	average	subway	line	station	highest	lowest	house price	sets
Middle-Class Issues	56	house	money	buy	bank	china	buy	white-collar	live	lords	house prices
South Yangtze	57	Nanjing	Plate	House Prices	Hexi	Yuan	Jiangbei	Wan	ASP	Buying	New
Tier-Market Analysis	58	city	frontline	line	china	population	future	beijing	housing price	property market	shanghai
Bride-price Story	59	Ji Bin	Wang Ting	Zhang Jiaying	No	Zhang	Marriage	Jiaying	TV	Buy	Shijiazhuang
Trade Log	60	square meters	deal	dollar	average price	million	set	residential	data	price	area

Cities Names	61	Guangzhou	Guangdong	Foshan	Region	Nanning	Nansha	Zhejiang	Province	Zengcheng	Places
Career Plan	62	Think	No	Job	Choice	Company	Many	Think	No	Things	Years
Interior Design	63	Design	Space	Inside	Out	Living	Home	Renovated	Furniture	Cave	New
School District	64	School	Elementary School	School District Housing	Education	Children	School District	Middle School	Experiment	Parents	Degrees
Professional Economic Analysis	65	House prices	China	Currency	Economy	Real estate	Growth	Growth	Inhabitants	High	Investment
Immigration	66	United States	China	Dollar	Million	Immigration	Chinese	New York	Graveyard	Lanzhou	Domestic
Financial Analysis	67	Banking	Finance	Investment	Loans	Capital	Real Estate	Assets	Risk	Finance	Stock Market
Capital-Area Cities	68	Beijing	Yanjiao	housing prices	Tongzhou	Beijing	Langfang	Gu'an	Wan	Tianjin	Central
Oversea Bubbles	69	Japan	China	Economy	Real estate	United States	Tokyo	Dollars	Bubbles	House prices	World
Oversea Properties	70	Finance	Phoenix	House Prices	Germany	Real Estate	Author	Copyright	Reply	Follow	Contact
Account Promotion	71	House Prices	Follow	Life	Introduction	Reply	Features	Information	Platform	Information	View
Life Story	72	Inside	Think	No	No	Go	Kind	Question	Times	Too	Eat
Market-Analysis-3	73	Market	Question	Price	No	Maybe	High	Very	Compare	Different	Two
Purchase News	74	For Sale	Price	Item	Land Market	Developer	Opening	News	New	Current	Home Buyer
New-District Plan	75	Wuhan	City	Development	Construction	Planning	Center	New District	New	Industry	Traffic
Purchase Limitation	76	Housing	Set	Loan	Policy	Provident Fund	Housing	Home	Purchase	Purchase	First
Vulgar Story-2	77	insurance	buy	aged	no	no	company	money	think	farmer	house prices
Property Tax	78	Realty Tax	Tax	China	Real Estate Tax	Real Estate	Express	House Prices	National	Imprompt	Real Estate
Investment Analysis	79	house price	buy	buy house	house	down	real estate	marketing	now	china	ten

**Appendix 1: The Entire 80 Topics and their Top 10 words (Translated in English)**

Topic Name	ID	Rank_0	Rank_1	Rank_2	Rank_3	Rank_4	Rank_5	Rank_6	Rank_7	Rank_8	Rank_9
Company Issues	0	公司	元	温州	万	企业	开发商	钱	资金	合同	家
Construction	1	面积	建筑	平方米	层	户型	工程	平米	住宅	总	高
China Development	2	中国	经济	发展	社会	未来	成为	企业	问题	国家	财富
Pearl 1st	3	深圳	房价	万	东莞	深圳市	龙华	惠州	元	片区	深
Pearl not-1st	4	元	均价	房	地址	珠海	大道	折	售	楼盘	交汇处
Price Rank	5	房价	排名	城市	名	市	全国	第名	均价	中国	县
House Agents	6	中介	二手房	价格	万	客户	业主	房子	小区	卖	房
Company News	7	地产	房地产	公司	房企	企业	项目	行业	万科	集团	元
Price Tease	8	房子	买	万	钱	卖	元	套	人民币	价格	约

Price Trends	9	房价	下跌	跌	楼市	上涨	市场	降价	涨	出现	已经
Travel Island	10	酒店	岛	吃	元	台湾	里	推荐	旅行	台北	沙滩
Price Complaint	11	媒体	买	降价	女人	套	成	没有	这种	政府	百姓
Market-Analysis-1	12	市场	城市	楼市	政策	房地产	房价	库存	上涨	需求	一线
Mortgage	13	银行	利率	贷款	房贷	元	利息	基准	还款	首	万
Used Price	14	元	均价	二手房	房价	上涨	上月	环	楼盘	新房	小区
Listing-1	15	花园	小区	国际	广场	新	公馆	城	家园	公寓	新村
not-1st Cities	16	郑州	品牌	河南	郑东	新区	郑州市	正	公司	消费者	市场
Positive Analysis	17	买	房子	房价	买房	房	套	现在	涨	没	想
Anhui	18	合肥	房价	元	滨湖	安徽	万	潍坊	幢	单元	价格
Increasing	19	房价	房地产	上涨	城市	房	中国	买房	发展	调控	房子
Service Industry	20	酒店	客人	服务	客户	费用	管理	产品	业务	使用	年月日
Negative Analysis	21	房子	涨	中国	跌	房地产	崩盘	国家	买	香港	现在
National Price	22	城市	价格	住宅	涨幅	上涨	房价	环比	下降	同比	新建
Mechanics Analysis	23	房价	城市	土地	收入	上海	上涨	涨	供应	北京	房子
Control Policy	24	政策	调控	楼市	限购	出台	市场	房地产	套	首	限贷
Life Plan	25	买	钱	工作	孩子	生活	没	两	父母	住	大学
Anecdote	26	最后	拨	潘	年代	石屹	没有	经济学	里	成	批
Procedure Supervise	27	登记	不动产	房地产	部门	销售	市场	预售	商品房	信息	价格
Headline Trash-News	28	中国	成为	北京	第一	世界	网络	没有	时代	悟空	吃
Market-Analysis-2	29	现在	中国	没有	市场	钱	增长	可能	问题	房子	经济
Contract Dispute	30	合同	房屋	违约	先生	双方	元	买卖	签订	要求	卖家
City Culture	31	北京	很多	中国	文化	少	全国	上海	男人	没	出
Other Price	32	翡翠	种	元	价值	万	玻璃	时间	罗瑛	大蒜	大学
Tier Price	33	城市	房价	上海	深圳	南京	涨幅	北京	广州	杭州	厦门
SE-Coastal Cities	34	厦门	房价	腾讯	房地产	房	福州	元	联合网	新闻	买房
South-west Cities	35	成都	房价	万	上海	天府	贵阳	圈	公里	条	板块
Vulgar Story-1	36	答	老师	张二狗	初中	狗	教育	省城	学习	没	里
Salary Expense	37	平均	工资	重庆	房价	中国	元月	吃	喝	北京	租金
Hong Kong	38	香港	平衡	内地	想	中国	次	块	房价	港元	美美
Travel Domestic	39	特色	元	三亚	大理	旅游	海南	丽江	西湖	住宿	活动
Specialist Analysis	40	房价	房地产	中国	泡沫	市场	经济	上涨	周期	政策	涨

Listing-2	41	元	平米	小区	花园元	花园	城元	均价	国际	房价	家园
Policy Announcement	42	发展	房地产	经济	市场	改革	政策	住房	工作	会议	中央
Money Supply	43	人民币	中国	美元	经济	汇率	贬值	美国	全球	资产	加息
North-China Cities	44	住宅	环比	下跌	价格	上涨	平方米	城市	指数	房价	均价
Oversea Market	45	房价	澳洲	悉尼	加拿大	房屋	房产	温哥华	万	地区	海外
Poor-Story-1	46	老人	中国	岁	卖	位	钱	买	出	孩子	没有
Land Finance	47	土地	政府	房地产	房价	中国	地方	问题	城市	财政	收入
Economic Analysis	48	中国	钱	房价	价格	经济	开发商	政府	没有	岛	交换
Land Bidding	49	苏州	地块	土地	元	地价	地王	万	宗	出让	楼面
East-China Cities	50	杭州	嘉兴	板块	浙江	城	萧山	滨江	钱江	楼盘	公寓
Urbanization	51	人口	城市	万	增长	地区	流入	房价	产业	占	发展
Poor-Story-2	52	杨佩	妈妈	孩子	没有	十字绣	绣	作品	喜欢	上海	朋友
Areal Service	53	区域	配套	地铁	线	中心	交通	项目	商业	规划	生活
Listing-3	54	元	小区	里	花园	家园	天津	新	公寓	楼盘	房价
Shanghai	55	上海	元	均价	地铁站	线	站	最高	最低	房价	套
Middle-Class Issues	56	房子	钱	买	银行	中国	买房	白领	住	地主	房价
South Yangtze	57	南京	板块	房价	河西	元	江北	万	均价	买房	新
Tier-Market Analysis	58	城市	一线	线	中国	人口	未来	北京	房价	楼市	上海
Bride-price Story	59	季彬	汪婷	张佳颖	没	张	结婚	佳颖	电视	买	石家庄
Trade Log	60	平方米	成交	元	均价	万	套	住宅	数据	房价	面积
Cities Names	61	广州	广东	佛山	地区	南宁	南沙	浙江	省	增城	地方
Career Plan	62	想	没有	工作	选择	公司	很多	觉得	没	事情	岁
Interior Design	63	设计	空间	里	出	生活	家	装修	家具	洞穴	新
School District	64	学校	小学	学区房	教育	孩子	学区	中学	实验	家长	学位
Professional Economic Analysis	65	房价	中国	货币	经济	房地产	增长	增速	居民	高	投资
Immigration	66	美国	中国	美元	万	移民	华人	纽约	墓地	兰州	国内
Financial Analysis	67	银行	金融	投资	贷款	资金	房地产	资产	风险	理财	股市
Capital-Area Cities	68	北京	燕郊	房价	通州	京	廊坊	固安	万	津	环
Oversea Bubbles	69	日本	中国	经济	房地产	美国	东京	美元	泡沫	房价	世界
Oversea Properties	70	财经	凤凰	房价	德国	房产	作者	版权	回复	关注	联系
Account Promotion	71	房价	关注	生活	介绍	回复	功能	信息	平台	资讯	查看
Life Story	72	里	想	没	没有	走	种	问题	次	太	吃

Market-Analysis-3	73	市场	问题	价格	没有	可能	高	非常	比较	不同	两
Purchase News	74	楼盘	价格	项目	楼市	开发商	开盘	记者	新	目前	购房者
New-District Plan	75	武汉	城市	发展	建设	规划	中心	新区	新	产业	交通
Purchase Limitation	76	住房	套	贷款	政策	公积金	房	家庭	购买	购房	首
Vulgar Story-2	77	保险	买	岁	没有	没	公司	钱	想	农民	房价
Property Tax	78	房产税	税	中国	房地产 税	房产	表示	房价	全国	征收	房地产
Investment Analysis	79	房价	买	买房	房子	跌	房产	营销	现在	中国	十

**Appendix 2: The Entire 80 Topics and their Top 10 words (in original Chinese)**